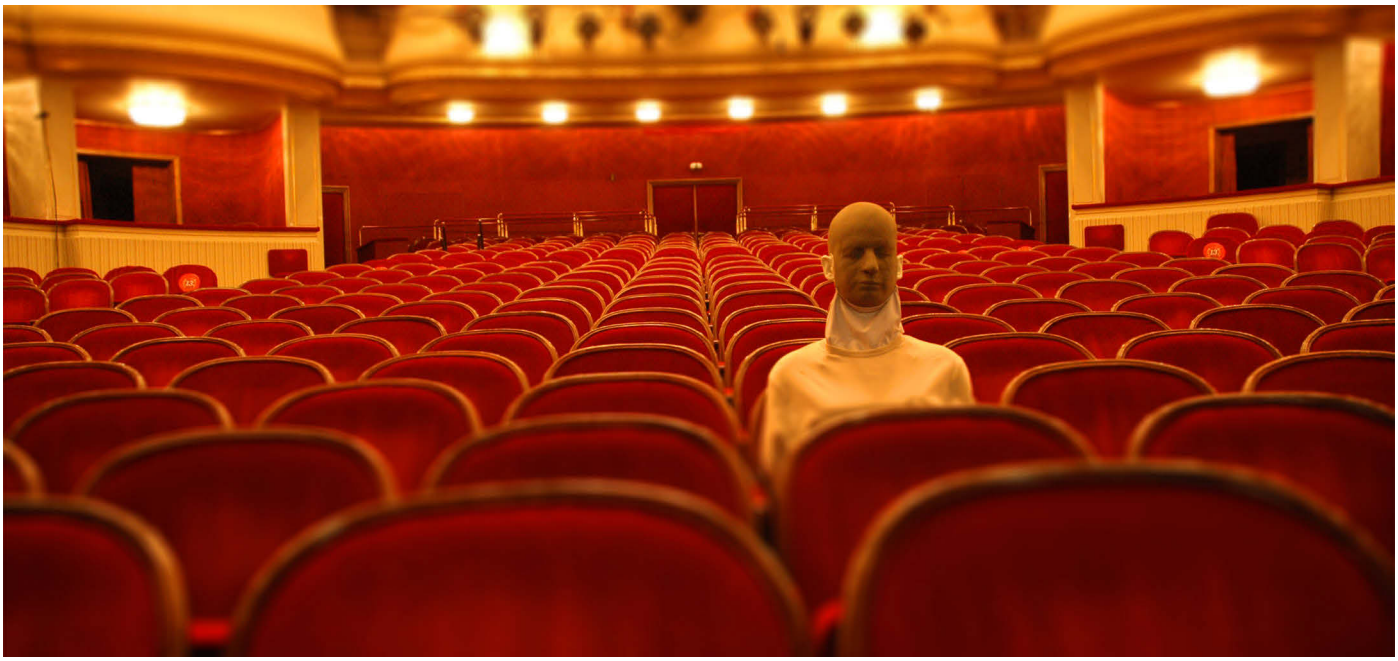
Stefan Weinzierl | Michael Vorländer | Franz Zotter | Hans-Joachim Maempel | Alexander Lindau (eds.)

Proceedings of the
EAA Joint Symposium on Auralization and Ambisonics
Berlin, 03th–05th April, 2014



ISBN 978-3-7983-2704-7 (online)

Universitätsverlag der TU Berlin

Technische Universität Berlin

Eds.: Stefan Weinzierl | Michael Vorländer | Franz Zotter |
Hans-Joachim Maempel | Alexander Lindau

Proceedings of the EAA Joint Symposium on Auralization and Ambisonics Berlin, 03th–05th April, 2014

Universitätsverlag der TU Berlin

Bibliographic information published by the Deutsche Nationalbibliothek
The Deutsche Nationalbibliothek lists this publication in the Deutsche Nationalbibliografie; detailed bibliographic data are available in the Internet at http://dnb.dnb.de.

Universitätsverlag der TU Berlin 2014
http://www.univerlag.tu-berlin.de

Fasanenstr. 88 (im VOLKSWAGEN-Haus), 10623 Berlin
Tel.: +49 (0)30 314 76131 / Fax: -76133
E-Mail: publikationen@ub.tu-berlin.de

Layout: Universitätsverlag der TU Berlin and Audio Communication Group, TU Berlin

ISBN 978-3-7983-2704-7 (online)

Online published on the Digital Repository of the Technische Universität Berlin:
URL http://opus4.kobv.de/opus4-tuberlin/frontdoor/index/index/docId/5414
URN urn:nbn:de:kobv:83-opus4-54140
http://nbn-resolving.de/urn:nbn:de:kobv:83-opus4-54140

Preface

In consideration of the remarkable intensity of research in the field Virtual Acoustics, including different areas such as sound field analysis and synthesis, spatial audio technologies, and room acoustical modeling and auralization, it seemed about time to organize a second international symposium following the model of the first EAA Auralization Symposium initiated in 2009 by the acoustics group of the former Helsinki University of Technology (now Aalto University). Additionally, research communities which are focused on different approaches to sound field synthesis such as Ambisonics or Wave Field Synthesis have, in the meantime, moved closer together by using increasingly consistent theoretical frameworks. Finally, the quality of virtual acoustic environments is often considered as a result of all processing stages mentioned above, increasing the need for discussions on consistent strategies for evaluation. Thus, it seemed appropriate to integrate two of the most relevant communities, i.e. to combine the *2nd International Auralization Symposium* with the *5th International Symposium on Ambisonics and Spherical Acoustics*. The Symposia on Ambisonics, initiated in 2009 by the Institute of Electronic Music and Acoustics of the University of Music and Performing Arts in Graz, were traditionally dedicated to problems of spherical sound field analysis and re-synthesis, strategies for the exchange of ambisonics-encoded audio material, and – more than other conferences in this area – the artistic application of spatial audio systems.

In the intense atmosphere of a single-track conference, more than 120 experts joined in Berlin for a three-day symposium. Thanks to the Stiftung Preußischer Kulturbesitz (SPK) the conference could be hosted in the inspiring environment of the Museum for Musical Instruments and the Federal Institute for Music Research, keeping one of the largest collections of musical instruments in Germany, and in the immediate vicinity of the Berlin Philharmonic Hall. The scientific program of the conference featured keynote lectures, regular and rapid poster talks, and a roundtable discussion on the concept of a planned Round Robin on Auralization. Moreover, several binaural and opto-acoustical room simulations as well as an immersive CAVE environment with photorealistic stereoscopic imaging and binaural rendering were presented to the audience. The program was completed by artistic contributions employing the 832 channel rectangular WFS array of the TU Berlin and the 26 channel hemispherical array of the Graz acoustics group for the presentation of live performances.

The publication at hand contains the official conference proceedings. It includes 29 manuscripts which have passed a 3-stage peer-review with a board of about 70 international reviewers involved in the process. Each contribution has already been published individually with a unique DOI on the *DepositOnce* digital repository of TU Berlin. Some conference contributions have been recommended for resubmission to *Acta Acustica united with Acustica*, to possibly appear in a Special Issue on Virtual Acoustics in late 2014. These are not published in this collection.

The preparatory work for this conference has largely been provided by a research consortium dedicated to the Simulation and Evaluation of Acoustic Environments (SEACEN, www.seacen.tu-berlin.de), installed in 2011 and funded by the German Research Foundation (DFG). Thus, the editors wish to thank the members of this consortium for their part in the organization of the conference and the publication of the scientific contributions, as they wish to thank all visitors of the conference for intense and inspiring discussions on each individual presentation!

Berlin, July 2014

The editorial board

Prof. Dr. Stefan Weinzierl
Audio Communication Group, TU Berlin
Coordinator SEACEN consortium

Prof. Dr. rer. nat. Michael Vorländer
Institute of Technical Acoustics, RWTH Aachen University
Co-Coordinator SEACEN consortium

Dr. rer. nat. Franz Zotter
Institute of Electronic Music and Acoustics, University of Music and Performing Arts, Graz

Dr. Hans-Joachim Maempel
Federal Institute for Music Research, Stiftung Preußischer Kulturbesitz

Dr. rer. nat. Alexander Lindau
Audio Communication Group, TU Berlin

Partners


European Acoustics Association
EAA


SEACEN


audiokommunikation
berlin


ITA
Lehrstuhl
für Technische
Akustik
RWTH AACHEN
UNIVERSITY


iem
kunst
uni
graz


Staatliches Institut für
Musikforschung
Preußischer Kulturbesitz


DEGA

Included Contributions

Proc. of the EAA Joint Symposium on Auralization and Ambisonics, Berlin, Germany, 3-5 April 2014

CHALLENGES IN THE CREATION OF ARTIFICIAL REVERBERATION FOR SOUND FIELD SYNTHESIS: EARLY REFLECTIONS AND ROOM MODES

Jens Ahrens

University of Technology Berlin
Ernst-Reuter-Platz 7
10587 Berlin, Germany
jens.ahrens@tu-berlin.de

ABSTRACT

Practical implementations of sound field synthesis evoke considerable artifacts that have to be considered in the creation of artificial reverberation. The most prominent artifact is spatial aliasing, which manifests itself as additional wave fronts that follow the desired synthetic wave front in time. These additional wave fronts propagate into different directions and occur at intervals that are similar to the intervals at which acoustic reflections occur in real rooms. It may be assumed that the human auditory system is not capable of differentiating aliasing artifacts and room reflections so that a synthetic reflection pattern should be designed such that it evokes a plausible pattern together with the aliased wave fronts. Two potential solutions are outlined. Finally, the capability of sound field synthesis of synthesizing room resonances (room modes) is analyzed and the promising results are illustrated based on numerical simulations.

1. INTRODUCTION

Sound field synthesis approaches employ high numbers of loudspeakers in order to synthesize a given desired sound field over an extended area [1]. The two best-known methods are Wave Field Synthesis (WFS) [2] and Near-field Compensated Higher Order Ambisonics (also termed Ambisonics with Distance Coding) [3]. The vast part of the scientific literature so far has focused on the synthesis of the direct sound of virtual sound sources. However, the creation of appropriate reverberation may be considered as important or even more important for achieving a desired spatial impression. Throughout the paper we assume the simple yet effective model of reverberation being composed of discrete early reflections the density of which increases over time and that gradually turn into diffuse late reverberation. The time interval after which the perceptual transition occurs is referred to as *mixing time* [4].

While the perceptual properties of mid-size and large rooms are mostly governed by the later part of the reverberation, small rooms can exhibit distinct early reflection patterns and low-frequency resonances also termed *room modes* [5, 6, 7]. This paper focuses on the creation of appropriate early reflection patterns as well as room modes. Late reverberation is not considered as solutions already exist as discussed below.

A first outline of the process of creating artificial reverberation for WFS can be found in [8] where a two-stage implementation is described. Early reflections are generated using a mirror image model [9] and late reverberation is generated using signals with appropriate statistical parameters. In [10] the capability of WFS of creating perceptually diffuse late reverberation via a set of plane waves is proven. Early reverberation was created using the mirror image model but was excluded from the evaluation. Appropriate input signals for the plane waves can be obtained, e.g., from microphones distributed in the recording venue as they can deliver sufficiently uncorrelated signals.

In [11] a convolution reverb is described that uses multipoint room impulse responses in order to create the proper reverberation for a given virtual sound source in WFS from dry (anechoic) source signals. Due to the large amount of data involved, a parameterization of the captured reverberation based on a plane wave representation and psychoacoustic criteria is proposed. However, no formal perceptual evaluation is provided. [12] presents an extension to the approach from [11] that enables the manipulation of measured multipoint impulse responses based on a three-dimensional visualization using augmented reality technologies. The manipulation is performed in time-frequency domain and its motivation is the provision of more flexibility and artistic freedom to the sound engineer.

None of the above mentioned approaches considers the mentioned artifacts that practical implementations of sound field synthesis exhibit. The synthesis of room modes has not been discussed in the literature. The present contribution discusses two approaches for the design of appropriate early reflection patterns considering the unavoidable spatial aliasing artifacts. It then investigates the potential of sound field synthesis of synthesizing room modes. When considering early reflections, we focus on artifacts as they appear in spatially fullband sound field synthesis methods such as WFS and the Spectral Division Method (SDM) [1, 13]. Spatially narrowband methods like the members of the Ambisonics family exhibit artifacts with slightly different properties. The extension of the presented results to narrowband methods will be outlined. The presented results on room modes are valid for all methods since room modes are only perceptually significant at lower frequencies [14, 15] where all methods exhibit a similar high accuracy [1].

This contribution focuses on the creation of artificial reverberation. It is not clear at this stage how the results can be transferred to reverberation recorded/measured with microphone arrays.

2. EARLY REFLECTIONS

2.1. Properties of the Spatial Aliasing Artifacts

The sound fields created by practical sound field synthesis systems exhibit a number of deviations from the prescribed virtual field [1]. These deviations are termed *artifacts* and the most important artifact in the present context is *spatial aliasing*. The term spatial

aliasing is typically used in a very broad sense and often refers to all artifacts that arise due to the combination of spatial discretization of the secondary source contour and the radiation properties of the involved secondary sources (i.e. the loudspeakers). Note that spatial aliasing can theoretically be avoided by using a continuous distribution of secondary sources. A detailed treatment can be found in [1, Sec. 4.4.4] and [16].

We illustrate the most relevant results based on the example scenario depicted in Fig. 1: A virtual plane wave that is synthesized by a circular distribution of 56 monopole loudspeakers. The properties of other non-focused virtual sources are very similar. Focused virtual sound sources are a special case in which pre-echoes arise [17]. They are excluded from the present investigation.

The two different basic options – spatially narrowband synthesis (27th order, Fig. 1(a)) and spatially fullband (infinite order, Fig. 1(b)) synthesis – are illustrated. It is evident that additional undesired wave fronts occur that follow the initial plane wave within a few ms (Fig. 2). A simple but useful model, especially for the spatially fullband example in Fig. 1(b), is the assumption that each active loudspeaker of the setup creates one additional spherical wave front that is emitted at the time instant at which the virtual plane wave passes the considered loudspeaker. As can be deduced from Fig. 2, the amplitudes of the additional wave fronts in Fig. 1(b) are only a few dB lower than the amplitude of the intended plane wave and are therefore perceptually relevant.

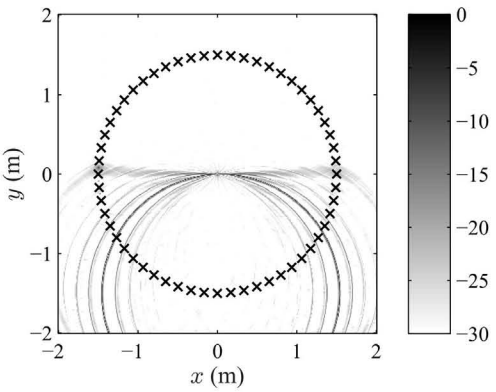
As illustrated in Fig. 1(c), the aliasing artifacts occur exclusively above a so-called aliasing frequency, which is approximately 1700 Hz in the current example[1]. Note that this behavior is very similar for spatially fullband and narrowband sound fields [1].

The timing and amplitude distribution of the aliasing artifacts is at least qualitatively similar to the timing and amplitudes of reflections in small reflective rooms. Fig. 3 shows some quantitative results. It is evident when comparing Fig. 3(a) and (c) that the artifacts arrive much denser than typical room reflections and in a time window that is much shorter. A situation in which a pattern evolves that is similar to the aliasing artifacts is when either the sound source or the receiver are located in a corner of the room. The proximity of the three walls that form the corner causes a very short delay between the direct sound and the first few reflections. Whether or not the two situations and spatial aliasing are similar from a perceptual point of view is not clear.

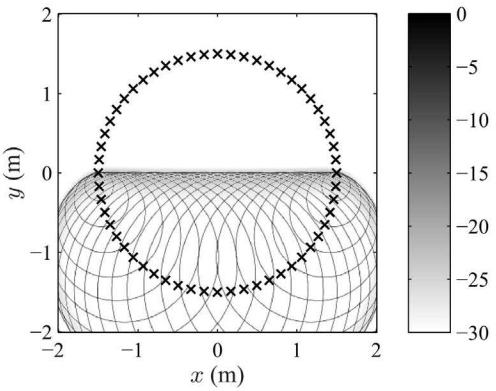
Another inconvenience arising from the densely spaced aliasing artifacts is the circumstance that the time interval between the direct sound – or the combination of direct sound and floor reflection – and the next following reflection is always very short in the artificial reverberation. Recording engineers often refer to this time interval as *pre-delay*. It can give important information about the size of the room and the location of the sound source. A large pre-delay suggests that the sound source is located at a significant distance from the closest wall. The room has therefore to be large. Manipulation of the pre-delay is a powerful audio mixing technique [18].

Despite certain differences, the working hypothesis in the present paper is that the human auditory system cannot distinguish between the aliasing artifacts and room reflections. This hypothesis bases mostly on the observations discussed above as well as on informal listening to setups like the one presented in [8], i.e. when the early room reflections are added as separate synthetic wave fronts (that exhibit their own aliasing artifacts). Informal listening shows that the reverberation in such a scenario tends to sound too dense. Note that each artificial reflection causes an entire set

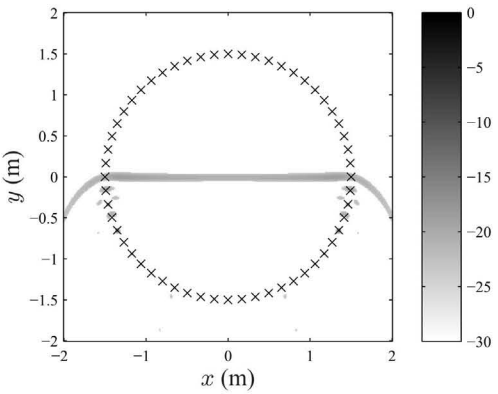
[1] A minimum-phase filter is applied in Fig. 1(c) in order to obtain a compact support of the resulting wave front.



(a) Spatially narrowband synthesis (e.g. Ambisonics).



(b) Spatially fullband synthesis (e.g. WFS, SDM).



(c) Sound field from Fig. 1(b) lowpass filtered with a minimum-phase FIR filter with a critical frequency of 1700 Hz.

Figure 1: Spatial impulse responses of a circular secondary source distribution in the horizontal plane when driven in order to synthesize a virtual plane wave propagating in positive y-direction. The absolute value of the time domain sound pressure is shown on in dB. (from [1, Fig. 4.19(c),(d)])

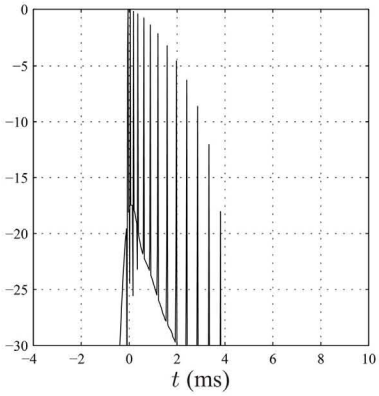


Figure 2: Time domain sound pressure of the field depicted in Fig. 1(b) at the coordinate origin on a logarithmic scale.

of wave fronts. A formal perceptual proof is not available at this point.

2.2. Adding the Low-frequency Content to the Aliased Wave Fronts

As noted above, a simplified interpretation of the aliasing artifacts is that an additional wave front arises from each loudspeaker. In the present context, we interpret these wave fronts as highpassed room reflections (recall Fig. 1(c)). The cutoff frequency (i.e. the spatial aliasing frequency) is typically between 1500 and 2000 Hz. It seems to be useful to artificially add the low frequency content to the highpassed reflections in order to make them natural. It should also be considered that room reflections experience diffraction at the boundaries of the reflecting surface at the very low end of the audible frequency range and a more modal behavior arises. It might therefore be preferable not to add the very low end to the (specular) reflection but treat it differently as discussed in Sec. 3. Note, however, that perception-based data are not available at this point.

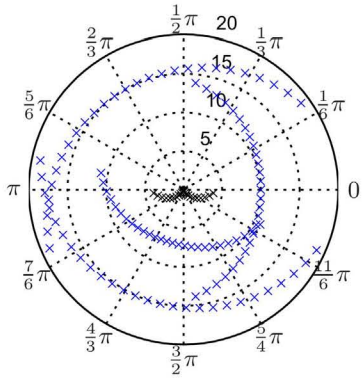
2.3. Adding Artificial Room Reflections

Adding artificial room reflections as separate synthetic wave fronts similar to the direct sound does not seem to be a favorable approach. Each added reflection will evoke a separate wave front pattern as illustrated by the blue marks in Fig. 3(a) so that only very few extra reflections lead to a very dense pattern of a high number of wave fronts. There are two obvious alternatives:

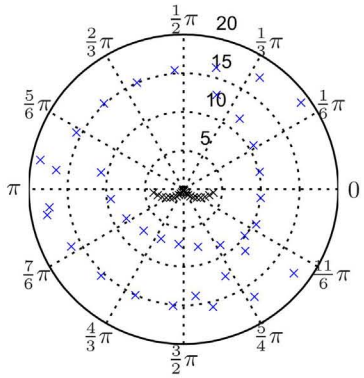
a) Use an individual loudspeaker for each reflection.

b) Synthesize the reflection using fewer loudspeakers with larger spacing than for the direct sound.

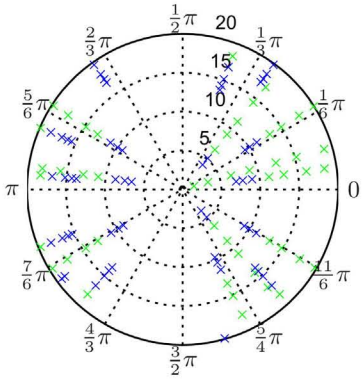
Ad a) This option is simple to implement and runs efficiently. However, the virtual room reflection that is produced by a single loudspeaker will exhibit an amplitude decay that is close to 6 dB for each doubling of the distance to the loudspeaker since the latter is typically very small and therefore acts like a monopole source. Larger sound sources – and therefore also reflections off large surfaces – exhibit a slower decay. This might not be an issue for small systems but with systems that have dimensions in the order of tens



(a) Timing and incidence azimuths of the wave fronts of the sound field from Fig. 1(b) at the coordinate origin (black marks) and four virtual reflections (blue marks).



(b) Fig. 3(a) but with the virtual reflections synthesized by every third loudspeaker.



(c) Timing and incidence azimuths of the first couple of reflections calculated from a mirror image model [9] for a room of dimensions $5 \times 3.1 \times 2.33$ m. Green marks: Reflections that impinge from more than 30° off the horizontal plane.

Figure 3: Timing and incidence azimuths of the wave fronts/reflections. The radius represents the delay in ms of a given wave front/reflection relative to the first arriving wave front.

of meters the relative amplitudes of the artificial reflections change differently with the listening location than real reflections.

Ad b) Using fewer loudspeakers with larger spacing causes aliased wave fronts that are less dense than those of the direct sound. It is therefore possible to achieve a more balanced distribution of the wave fronts than when all loudspeakers are used as depicted in Fig. 3(b). Using more loudspeakers causes a slightly slower amplitude decay of the wave fronts than for a)[2].

Note that for both options a) and b) the incidence angles of the virtual reflections change somewhat differently with the listening location than those of real reflections. The perceptual significance of this circumstances is not clear.

2.4. Extension to Spatially Narrowband Methods

As mentioned in the Introduction, spatially narrowband methods for sound field synthesis such as the Ambisonics family of approaches exhibit spatial discretization artifacts that have somewhat different properties than those of the spatially fullband methods discussed so far. As evident from comparing Fig. 1(a) and (b), the additional wave fronts are fewer and are less homogeneously distributed over the listening area. Note that this is more of a quantitative than a qualitative difference. We assume that the considerations presented in the previous sections hold.

3. ROOM MODES

3.1. Physical Fundamentals

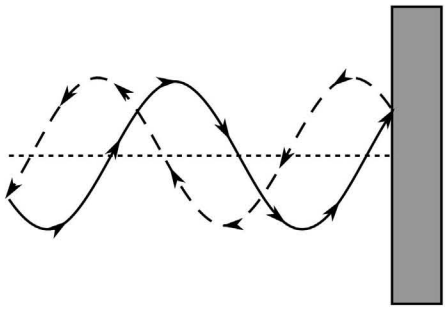
In order to illustrate the physical circumstances that lead to room modes, i.e., resonances of the room, we use the simplified model of a plane wave bouncing off an infinite rigid plane that extends normal to the propagation direction of the plane wave. We neglect phenomena like diffraction that occur at boundaries of finite extent. The reader interested in a detailed treatment is referred to [7].

Rigid boundaries such as the walls of a room constitute Neumann boundary conditions for the sound waves inside the rooms as the particle velocity in the propagation direction of the wave vanishes at the boundary (the particles cannot move due to the boundary). As a consequence, the wave bounces back without a phase shift of the sound pressure. Refer to Fig. 4(a) for an illustration. The result is a field that consists of two plane waves of equal frequency and amplitude but with opposing propagation directions. Expressed in one-dimensional and in complex notation, this reads

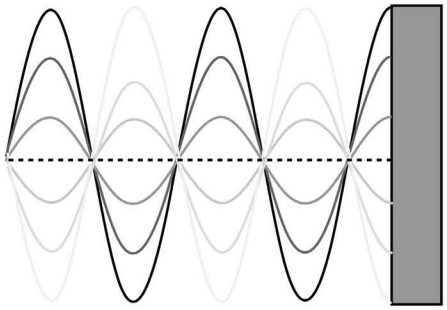
$$p_{\text{incoming}} + p_{\text{reflected}} = \mathrm{e}^{-\mathrm{i}\frac{\omega}{c}x}\mathrm{e}^{\mathrm{i}\omega t} + \mathrm{e}^{\mathrm{i}\frac{\omega}{c}x}\mathrm{e}^{\mathrm{i}\omega t} = 2\cos\left(\frac{\omega}{c}x\right)\mathrm{e}^{\mathrm{i}\omega t} \tag{1}$$

when assuming monochromatic waves of unit amplitude and choosing the time reference and coordinate system such that the two waves exhibit a relative phase shift of 0. The result is a standing wave of equal frequency like the component waves and with a pressure antinode at the boundary as illustrated in Fig. 4(b).

Note that this formation of standing waves occurs at all frequencies. We are usually not dealing with monochromatic waves but with waves that carry broadband and therefore time-varying content. In real rooms the waves bounce back and forth between the bounding surfaces. Standing waves occur only at particular frequencies in a room, namely at those frequencies for which the path length of the periodic path that a wave travels inside the room corresponds to an integer multiple of half the wavelength after the system has reached a steady state. The simplest case is a wave bouncing between two parallel walls of infinite extent. When the propagation direction of the wave is perpendicular to the walls and when the walls are perfectly reflective a persistent standing wave evolves at those frequencies specified above. Depending on the periodic path, different standing wave pattern evolve.



(a) One-dimensional wave impinging on a rigid boundary (solid line) and being reflected off (dashed line).



(b) Standing wave resulting from a superposition of the impinging and the reflected wave. Different shading of the wave refers to different time instants.

Figure 4: Illustration of the formation of a standing wave due to reflection off a rigid surface.

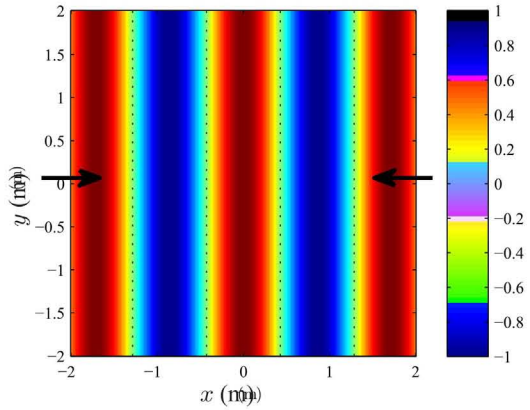
A significant amount of diffraction occurs in real rooms especially at low frequencies where the wave length is of similar order like the dimensions of the wall so that always a wave component that propagates perpendicular to a given wall arises. The amplitude and the Q-factor (and therefore the ringing duration) of the resonance depend on the acoustical properties of the boundaries, which are usually not perfectly rigid. Room modes occur all over the audible frequency bandwidth but only the low-frequency modes are perceptually relevant because of their sparsity [14, 15].

Fig. 5 illustrates the node/antinode patterns that evolve for the combination of different numbers of plane wave pairs so that different patterns can be realized. The standing waves exhibit their maximum amplitudes at the depicted time instant. Refer also to the animations at [19] that accompany this paper.

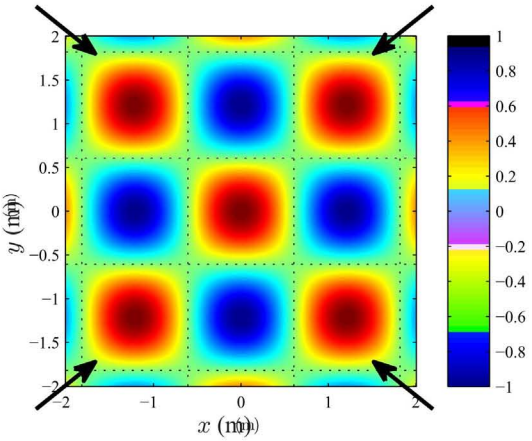
3.2. Room Modes in Sound Field Synthesis

The parameters of modes in real rooms are complicated to determine because they depend heavily on the position of the source and many of the acoustical properties of the room. The interested reader is referred to [7]. It may be doubted that the human auditory system has a detailed expectation of plausible room modes so

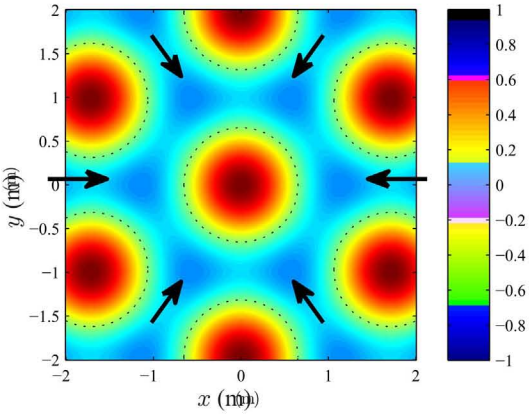
[2] Note that the slowest amplitude decay that e.g. an infinite linear loudspeaker array can produce is that of a cylindrical wave of 3 dB attenuation per doubling of the distance [1].



(a) 1 pair of plane waves.



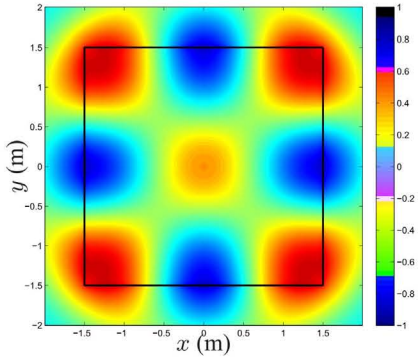
(b) 2 pairs of plane waves.



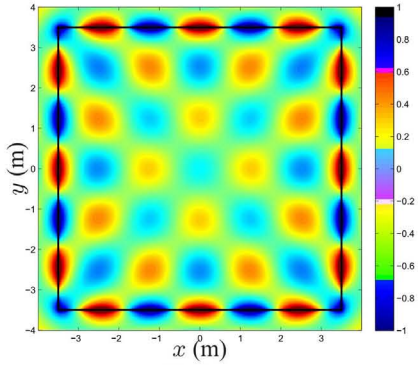
(c) 3 pairs of plane waves.

Figure 5: Cross-sections through the horizontal plane of the sound pressure evolving from different numbers of plane wave pairs of frequency $f = 200$ Hz. All sound fields are normalized to unity. The arrows illustrate the propagation directions of the plane wave components. The dotted lines mark the nodes.

that the more pragmatic approach of choosing the parameters like resonance frequency, amplitude, and bandwidth based on simple



(a) 3×3 m



(b) 7×7 m

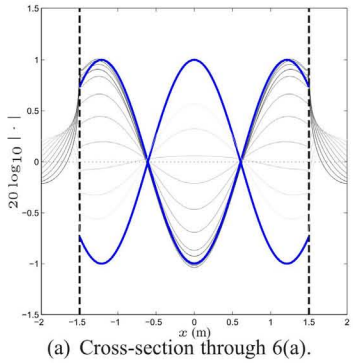
Figure 6: The standing wave from Fig. 5(b) synthesized by two quadratic sound field synthesis systems of different sizes. The black lines represent the secondary source contours. Note the different scalings of the axes.

statistical assumptions might be sufficient.

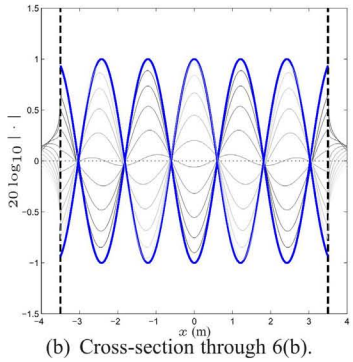
The creation of room modes based on pairs of plane waves as described in Sec. 3.1 is straightforward in three-dimensional sound field synthesis systems such as spherical arrangements of loudspeakers because plane waves with low-frequency content can be synthesized accurately. A set of narrow peak filters can be applied to the input signal of a given virtual source to efficiently create the narrowband input signals for the plane wave pairs.

The situation is more challenging in 2.5-dimensional – i.e. horizontal-only – synthesis. Here, synthetic plane waves exhibit an unavoidable amplitude decay of 3 dB for each doubling of the distance to the loudspeakers [1]. Short arrays exhibit an even faster amplitude decay because of the spatial truncation. Fig. 6 depicts the sound field from Fig. 5(b) synthesized by two loudspeaker systems of different sizes. Fig. 7 shows cross-sections through Fig. 6(a) and (b). Refer also to the animations at [19].

The simulations from Fig. 6 and 7 indicate that it is indeed possible to achieve standing waves in 2.5D synthesis. The deviation of the synthesized sound wave from the theoretic standing wave is small even for the mid-size array with 7 m edge length. It seems that the propagating components in the synthesized sound field are negligible. Future work has to investigate in what situations a considerable perceptual impairment arises.



(a) Cross-section through 6(a).



(b) Cross-section through 6(b).

Figure 7: Cross-section through Fig. 6(a) and (b) along the x-axis. Different gray shading represents different time instants. The blue lines represent the envelope of the prescribed (exact) standing wave calculated similarly to (1).

4. CONCLUSIONS

We presented guidelines for the creation of artificial early reflections and room modes in sound field synthesis. The most important aspect in the context of the creating of early reflections is the fact that practical sound field synthesis systems exhibit artifacts that are known as spatial aliasing that exhibit properties that are similar to room reflections. We presented two approaches for the design of reflection patterns that take the spatial aliasing artifacts into account.

We also suggested that the modal behavior of real rooms can be mimicked by synthesizing pairs of plane waves that propagate into opposing directions. Numerical simulations showed promising results even for 2.5D synthesis where the synthesized plane waves exhibit an undesired amplitude decay.

Acknowledgments

The author thanks Frank Schultz for valuable comments on the manuscript. The work presented in this paper is supported by grant AH 269/2-1 of Deutsche Forschungsgemeinschaft.

5. REFERENCES

[1] J. Ahrens, *Analytic Methods of Sound Field Synthesis*, Springer-Verlag, Berlin/Heidelberg, 2012.

[2] A. J. Berkhout, D. de Vries, and P. Vogel, "Acoustic control by wave field synthesis," *JASA*, vol. 93, no. 5, pp. 2764–2778, May 1993.

[3] J. Daniel, "Spatial sound encoding including near field effect: Introducing distance coding filters and a viable, new Ambisonic format," in *23rd International Conference of the AES*, Copenhagen, Denmark, May 2003.

[4] A. Lindau, L. Kosanke, and S. Weinzierl, "Perceptual evaluation of model- and signal-based predictors of the mixing time in binaural room impulse responses," *JAES*, vol. 60, no. 11, pp. 887–898, Nov. 2012.

[5] J. Blauert and W. Lindemann, "Auditory spaciousness: Some further psychoacoustic analyses," *JASA*, vol. 80, no. 2, pp. 533–542, 1986.

[6] H. Kuttruff, *Room Acoustics*, Spon Press, London, fifth edition edition, 2009.

[7] F. Mechel, *Room Acoustical Fields*, Springer-Verlag, Berlin/Heidelberg, 2013.

[8] D. de Vries, A. J. Reijnen, and M. A. Schonewille, "The wave field synthesis concept applied to generation of reflections and reverberation," in *96th Convention of the AES*, Amsterdam, The Netherlands, 1994.

[9] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *J. Acoust. Soc. Am*, vol. 65, no. 4, pp. 943–948, Apr. 1979.

[10] J.-J. Sonke, "Variable acoustics by wave field synthesis," PhD thesis, Delft University of Technology, 2000.

[11] E. Hulsebos, "Auralization using wave field synthesis," PhD thesis, Delft University of Technology, 2004.

[12] F. Melchior, "Investigations on spatial sound design based on measured room impulses," PhD thesis, Delft University of Technology, 2011.

[13] J. Ahrens and S. Spors, "Sound field reproduction using planar and linear arrays of loudspeakers," *IEEE Trans. on Sp. and Audio Proc.*, vol. 18, no. 8, pp. 2038–2050, Nov. 2010.

[14] J. Fazenda, "Perception of room modes in critical listening spaces," PhD thesis, University of Salford, 2004.

[15] M. Karjalainen, P. Antsalol, A. Mäkivirta, and V. Välimäki, "Perception of temporal decay of low frequency room modes," in *116th Convention of the AES*, Berlin, Germany, May 2004.

[16] S. Spors and J. Ahrens, "Spatial aliasing artifacts of wave field synthesis for the reproduction of virtual point sources," in *126th Convention of the AES*, Munich, Germany, May 2009.

[17] S. Spors, H. Wierstorf, M. Geier, and J. Ahrens, "Physical and perceptual properties of focused sources in wave field synthesis," in *127th Convention of the AES*, New York, NY, Oct. 2009, p. paper 7914.

[18] R. Izhaki, *Mixing Audio - Concepts, Practices and Tools*, Focal Press, Oxford, 2007.

[19] J. Ahrens, "Animations," `http://www.soundfieldsynthesis.org/animations/eaa-2014`.

PERCEPTUAL OPTIMIZATION OF ROOM-IN-ROOM REPRODUCTION WITH SPATIALLY DISTRIBUTED LOUDSPEAKERS

Julian Grosse

Cluster of Excellence "Hearing4all"
Acoustics Group
Carl von Ossietzky University
Oldenburg, Germany
julian.grosse@uni-oldenburg.de

Steven van de Par

Cluster of Excellence "Hearing4all"
Acoustics Group
Carl von Ossietzky University
Oldenburg, Germany
steven.van.de.par@uni-oldenburg.de

ABSTRACT

It is often desirable to reproduce a specific room-acoustic scene, e.g. a concert hall in a playback room, in such a way that the listener has a plausible and authentic spatial impression of the original sound source including the room acoustical properties. In this study a perceptually motivated approach for spatial audio reproduction is developed. This approach optimizes the spatial and monaural cues of the direct and reverberant sound separately. More specifically, the (monaural) spectral cues responsible for the timbre and the (binaural) interaural cross correlation (IACC) cues, responsible for the listener envelopment, were optimized in the playback room to restore the auditory impression of the recording room. The direct sound recorded close to the source is processed with an auditory motivated gammatone filterbank such that the spectral cues, ITD's and ILD's are comparable to the direct sound in the recording room. Additionally, the reverberant sound, which was recorded at two distant locations from the source, is played back via dipole loudspeakers. Due to the arrangement of the two dipole loudspeakers, only the diffuse sound field in the playback room is excited, therefore the spectral cues and the IACC of the reverberant sound field can be controlled independently to match the cues that were present in the recording room. As indicated by a preliminary listening test the applied optimization is perceptually similar to the reference signal and is generally preferred when compared to a conventional room-in-room reproduction.

1. INTRODUCTION

The perception of a sound source strongly depends on the room (e.g. church or a concert hall) in which the source is placed. When a recorded sound source is reproduced, it is desirable to not only faithfully reproduce the sound source but also the room acoustics of the recording room. There are several methods to reproduce a sound field which are based on using large arrays of loudspeakers, e.g. Ambisonics [1] or Wave-field-synthesis (WFS) [2]. Since the above-named methods need a large number of loudspeakers, they are less suitable for sound reproduction in the living room. Furthermore, these approaches assume that the room where the loudspeaker array is placed has no boundaries and the propagating sound wave is not affected by the room. The inaccuracies that will occur due to a reverberative environment are generally not considered.

Some problems that will occur when a sound recorded in a 'recording' room is reproduced in another echoic 'reproduction' room can be understood when considering that in this case that the listener who is present in the reproduction room will effectively hear the combined room acoustics of both rooms. This implies that the Room Impulse Responses (RIR) of both rooms are convolved with one another. As a consequence, the envelope of the resulting impulse response will look like a second order system. Additionally, also the spectral statistics will change. For a single RIR the standard deviation in the magnitude spectrum is approximately $\sigma = 5.5$ dB [3]. Due to the convolution of both Room Impulse Responses the standard deviation of the magnitude spectrum will increase by a factor of $\sqrt{2}$ which may be perceived as an increase in spectral coloration.

This study presents a method that compensates for the detrimental effects of the reproduction room using human auditory perceptual criteria. Thus this method will not attempt to reproduce the sound in an exact physical way at the eardrum of a listener. Instead it optimizes timbre and spatial characteristics based on auditorily motivated frequency bands and on the interaural cross correlation. For optimization an artificial head is placed in the playback room so that the spatial and timbre cues can be matched to a reference artificial head in the recording room. In normal loudspeaker playback, loudspeakers are designed such that the direct sound path has a flat transfer function. As a consequence, there is little control over how the reverberant sound field in the playback room is excited. In our approach, a set of rear dipole loudspeakers will be used the excite the reverberant sound field separately, by aligning the dipole loudspeakers such that the listener receives no direct sound path from the dipole loudspeakers. In this way both the timbre and the spatial properties of the reverberant sound field in the playback room can be controlled separately. This approach has the restriction that the reproduction room needs a smaller reverberation time than the recording room.

Similar perceptually motivated approaches for sound reproduction have been investigated before. De Bruijn et al. [4] showed with a similar setup that it is possible to modify the perceived distance by separately presenting the direct sound over a conventional stereo loudspeaker setup and the reverberant sound over rear loudspeaker. Breebaart et al. [5] found in the context of low-bit-rate audio coding that the binaural attributes ILD (interaural level differences), ITD (interaural time differences) and the IACC (Interaural Cross Correlation) can sufficiently describing the spatial percept of a stereo audio signal. The ILD's and ITD's are responsible for localization and the IACC is responsible for the perceived spaciousness and the listener envelopment [6].

In this study in order to be able to optimize the direct sound and the reverberant sound field separately, the direct sound is recorded close to the sound source in the recording room and is presented

over a stereo loudspeaker setup. The reverberant sound is recorded at two distant positions and is rendered via two dipole loudspeakers. This loudspeaker arrangement gives the possibility to control the overall timbre of the reproduced sound source, the IACC, and the reverberation time of the recording room including the effect of the rendering in the playback room.
In order to evaluate the authenticity of the sound reproduction that can be obtained with the proposed method, the reference signal is compared with the applied optimization over headphones in a MUSHRA-Test (MUltiple Stimulus test with Hidden Reference and Anchor). For additional simulated listening situations, a multichannel reproduction and a more conventional room-in-room reproduction method is used.

2. METHOD

In this section the optimization method is described in detail. An artificial head is placed in the recording room (as the reference) as well as in the playback room (Fig.1) which are used for recording binaural room impulse responses (BRIRs). In this approach, the optimization does not aim for an accurate reproduction of the physical sound field at the ear-drum of the artificial head, but rather the excitation pattern created on the basilar membrane is considered [7]. In this context, the direct and the diffuse sound are optimized separately taking into account the room-acoustical properties of the playback room. In Section 2.1 it will be described how the BRIR is divided into a direct path and a diffuse sound path. In Section 2.2 the analysis of the playback room is described. In Section 2.5 the method is introduced for optimizing the reproduction in the playback room based on perceptual criteria measured at the "ear-drum" of the artificial head.

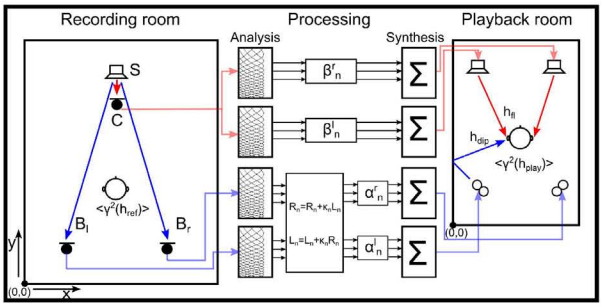


Figure 1: *Left: Recording room with a reference artificial head, the close microphone $C(t)$ for the direct sound and two omnidirectional microphones $B_{l,r}(t)$ in the diffuse sound field. Right: The playback room with the artificial head for the perceptual optimization, two front loudspeakers (60° stereo triangle) and the dipole loudspeakers to excite the diffuse sound field. The processing-stage is divided into an analysis and a synthesis stage. In the analysis-stage the energy outputs of the gammatone filters are observed. Each filter output is multiplied with a real-valued gain factor (β_n for the direct sound $C(t)$, α_n for the reverberant sound $B(t)$) to control the overall coloration. The mixing factor κ_n controls the IACC in the playback room. The synthesis stage includes the phase-alignment of the gammatone filters and finally sums up all sub-bands.*

2.1. Analysis - Recording room

In this section the analysis of two parameters, the timbre and the cross correlation is described. The first parameter is the timbre expressed in terms of the excitation pattern determined for the recording room. The analysis of the the recording room will occur for the direct sound and for the reverberant sound separately to optimize the timbre in the playback room. To clarify the notation conventions, only the left (noted as l) ear of the artificial head is considered in the following. Derivations for the right ear are similar and only require l to be replaced by r and vice versa. If we observe a BRIR in the recording room $h(t)_{ref}$ we can split the BRIR (denoted with the subscript ref) into two parts.

$$h(t)^{(l)}_{ref} = h(t)^{(l)}_{d,ref} + h(t)^{(l)}_{rev,ref} \quad (1)$$

where the indices d, rev are indicating the direct sound and the reverberative sound in the recording room, respectively. The separation of the two parts will occur with the separation time constant t_m. For the separation, a squared cosine window is used with a 4 ms flank. The separation time constant t_m in the optimization will effectively control the direct to diffuse ratio, therefore it controls the T_{60} reverberation time in the playback room. The derivation of the optimal separation time constant can be found in section 3.1. After separating the BRIR, the two parts are filtered with an auditory motivated 4^{th}-order gammatone filterbank (GTFB) (cf. [8]). The filters are distributed equally on an ERB scale (equivalent rectangular bandwidth) in the range of 20 Hz to 24 kHz. This yields 42 gammatone channels. The filtered BRIR signal h in each gammatone channel n is denoted by $\langle\gamma_n(h)\rangle$. The overall excitation pattern is determined by:

$$\langle\gamma_n^2(h(t)^{(l)}_{d,ref})\rangle = \int_{t_d}^{t_m}\int_{-\infty}^{+\infty} |h(\tau)^{(l)}_{ref} * \gamma_n(t-\tau)|^2 d\tau \quad (2)$$

where t_d indicates the start of the impulse response. The excitation pattern of the reverberant part is calculated by integrating the BRIR from t_m until the end of the BRIR. The expression $\langle\gamma_n^2(h(t))\rangle$ contains the energy in each frequency band n at the center frequency f_c of the gammatone filters (called excitation pattern). This excitation pattern control the overall energy in the playback room.
The second parameter analysed is the interaural cross correlation (IACC) in the recording room. The normalized cross correlation coefficient of the whole BRIR is determined by observing the time-signal in each gammatone channel n for the whole BRIR. The IACC is processed in the following way:

$$IACC[q,n] = \frac{\sum_{m=-\infty}^{\infty} h[m,n]_{ref,l} \cdot h[m+q,n]_{ref,r}}{\sqrt{(\sum_{m=-\infty}^{\infty} h[n]^2_{ref,l})} \cdot \sqrt{(\sum_{m=-\infty}^{\infty} h[n]^2_{ref,r})}} \quad (3)$$

in which l and r are the left and right channels of the artificial head and q is the time delay in samples. Within this context the value at $q = 0$ is used.

2.2. Analysis - Playback room

The analysis of the playback room is quite similar to the analysis of the recording room. The complete BRIR in the playback room (denoted with the subscript $play$) is defined by:

$$h(t)^{(l)}_{play} = h(t)^{(l)}_{d,play} + h(t)^{(l)}_{rev,play} \quad (4)$$

The BRIR can again be separated into a direct and a reverberant part. The separation time constant is the same as in Section 2.1. h_{pr} is the BRIR in the playback room (room-in-room (RinR)) when rendering the impulse response measured with microphone C and B. The correction factor β_n is used for the direct sound

in each band and α_n is needed to control the overall energy, thus, the amount of the diffuse field in the playback room. Theoretically, since the recording microphone, $C(t)$, is close to the sound source, only direct sound is recorded, and the RIR is a simple convolution with the BRIR of the loudspeaker to the artificial head in the playback room. We will, however, consider the more general case where $C(t)$ also incorporates some reverberation. In addition the diffuse sound field is excited separately with dipole loudspeakers:

$$h(t)^{(l)}_{pr} = \beta^{(l)}_n [C(t) * h(t)^{(l)}_{play}] + \alpha^{(l)}_n [(B(t)^{(l)} * h(t)^{(ll)}_{dip}) + (B(t)^{(r)} * h(t)^{(rl)}_{dip})] \quad (5)$$

where the superscript ll refers to the path from the left loudspeaker to the left ear and the superscript rl refers to the path from the right loudspeaker to the left ear of the artificial head. The BRIR $h(t)_{dip}$ of the dipole loudspeakers are convolved with the signals B which were recorded at two distant positions in the recording room. Due to the directivity pattern of a dipole loudspeaker it is possible to excite only the diffuse sound field in the playback room when the zero is directed towards the listener. Because we know that h_{play} can be separated into a direct and a diffuse part we can express Equation 5 as:

$$h(t)^{(l)}_{pr} = \beta^{(l)}_n \left[C(t) * h(t)^{(l)}_{d,play} + C(t) * h(t)^{(l)}_{rev,play} \right] + \alpha^{(l)}_n \left[(B(t)^{(l)} * h(t)^{(ll)}_{dip}) + (B(t)^{(r)} * h(t)^{(rl)}_{dip}) \right] \quad (6)$$

The aim is to make the BRIR measured in the recording and playback rooms equal:

$$h(t)^{(l)}_{ref} = h(t)^{(l)}_{pr} \quad (7)$$

One classical approach would be to equalize the transfer function such that the sound pressure signal at the listeners eardrum is the same in both rooms (like crosstalk cancellation). In the perceptive approach of this study, however, we do not want to optimize the transfer function in an exact physical sense but rather in a physiological sense, i.e. by optimizing the levels measured at the output of the auditory filters. Thus, we only consider the excitation pattern. We can express Equation 7 as:

$$\langle \gamma^2_n (h(t)^{(l)}_{ref}) \rangle = \langle \gamma^2_n (h(t)^{(l)}_{pr}) \rangle \quad (8)$$

Now we can substitute all given impulse responses of the recording and the playback room. By solving Equation 8, the cross terms between the direct sound field and the diffuse sound field are cancelled out because of the assumption that the direct signal are incoherent to the diffuse signal. The resulting final term is shown in Equation 9.

$$\underbrace{\langle \gamma^2_n (h(t)^{(l)}_{d,ref}) \rangle}_{p1} - \beta^2_n \cdot \underbrace{\langle \gamma^2_n (C(t) * (h(t)^{ll}_{d,play} + h(t)^{(rl)}_{d,play})) \rangle}_{p2} + \underbrace{\langle \gamma^2_n (h(t)^{(l)}_{rev,ref}) \rangle}_{p3} - \underbrace{\beta^2_n \cdot \langle \gamma^2_n (C(t) * (h(t)^{(ll)}_{rev,play} + h(t)^{(rl)}_{rev,play})) \rangle}_{p4} - \underbrace{\alpha^2_n \cdot \langle \gamma^2_n ((B(t)^{(l)} * h(t)^{(ll)}_{dip}) + (B(t)^{(r)} * h(t)^{(rl)}_{dip})) \rangle}_{p5} \stackrel{!}{=} 0 \quad (9)$$

Because the direct sound in the playback room also affects the excited diffuse field in this room (noted in Eq. 9 as $p4$), the direct sound has to be adjusted first. To match the direct sound in the playback room to the reference, the factor β^2_n in Equation 10 has to be processed. The term $p4$ in Eq. 9 where a β^2_n appears does not appear in this equation because the adjusted expression is taken into account when α^2_n is calculated. Equation 10 contains the energy of the direct sound in the recording room and the playback room.

$$\beta^2_{n,l} = \frac{\langle \gamma^2_n (h(t)^{(l)}_{d,ref}) \rangle}{\langle \gamma^2_n (C(t) * (h(t)^{(ll)}_{d,play} + h(t)^{(rl)}_{d,play})) \rangle} \quad (10)$$

The direct sound adjustment makes sure that the energy of direct sound is comparable in both rooms. The overall timbre can now be controlled via the dipole loudspeakers. In Equation 9 the difference between the first two terms $p1$ and $p2$ should be zero because the energy of the direct sound in the playback room was adjusted one step before. This leads in Equation 9 to the following expression for the diffuse sound field:

$$\alpha^2_{n,l} = \frac{\langle \gamma^2_n (h(t)^{(l)}_{rev,ref}) \rangle}{\langle \gamma^2_n (B(t)^{(l)} * h(t)^{(ll)}_{dip}) + (B(t)^{(r)} * h(t)^{(rl)}_{dip}) \rangle} - \frac{\beta^2_{n,l} \langle \gamma^2_n (C(t) * (h(t)^{(ll)}_{rev,play} + h(t)^{(rl)}_{rev,play})) \rangle}{\langle \gamma^2_n (B(t)^{(l)} * h(t)^{(ll)}_{dip}) + (B(t)^{(r)} * h(t)^{(rl)}_{dip}) \rangle} \quad (11)$$

An interesting property of Equation 9 is that the term $\langle \gamma^2_n (C(t) * h(t)_{rev,play}) \rangle$ describes the diffuse part of the BRIR in the playback room which is excited by the front loudspeakers. The Equations 10 and 11 can be solved by adapting the excitation pattern of the particular part in the playback room to the excitation pattern in the recording room.

2.3. Coefficient processing

For solving the values of alpha and beta, only the magnitude response is considered. Equation 12 shows exemplarily how the matrix A is computed for the direct path in the playback room.

$$A_{n,f} = |\sum_{n=1}^{P} \gamma_n(f) \cdot H(f)_{d,play}|^2 \quad (12)$$

where each row corresponds to the magnitude transfer function of the gammatone filtered signal. Equation 13 shows the transfer function of the direct path in the recording room.

$$b = |H(f)_{d,ref}|^2 \quad (13)$$

In Equation 14, A is a matrix (known) and α^2 (unknown) and b (known) are vectors.

$$A \cdot \alpha^2 = b \quad (14)$$

If the Matrix A has more rows than columns, the simple solution $\alpha = A^{-1} \cdot b$ can not be applied because A is not a square matrix. In our case we do not have a square Matrix and so we have a overdetermined problem which can be solved using the method of least squares:

$$\alpha^2 = (A^H \cdot A)^{-1} \cdot A^H \cdot b \quad (15)$$

at which superscript H resembles the conjugate transposition of the matrix A. The solution α gives us the gain-factors for each bandpass and can be multiplied in the frequency or time-domain by taking the square root of each element in α^2. In our specific case, the

vector α^2 is the wanted coefficient α_n^2 for the dipole loudspeakers and β_n^2 for the stereo loudspeakers. This solution was suggested by [9].

2.4. IACC optimization

The next step is to optimize the IACC. The correlation strongly depends on the optimization of the direct and diffuse sound. Therefore, the optimization of the IACC is done iteratively by mixing the signals of the omnidirectional microphones in the following way:

$$B_n^{l'} = B_n^l + \kappa_n \cdot B_n^r \tag{16}$$

$$B_n^{r'} = B_n^r + \kappa_n \cdot B_n^l \tag{17}$$

were κ_n is varied iteratively in the range of [-1:1] with a step size of 0.2 in each band n to control the IACC via the dipole loudspeakers. If we apply a $\kappa_n = 1$, the omnidirectional microphone signals B have a maximum correlation. With $\kappa_n = 0$ the signals are mostly decorrelated.

- Step 1: Adjust direct sound such that it is comparable to the direct sound in the recording room (according to Equation 10).
- Step 2: Mix the omnidirectional microphone signals (according to Equation 16 and 17) in the range of [-1:1] in each frequency band n.
- Step 3: Optimize the dipole loudspeaker signals according to Equation 11 that the overall energy in the playback room is comparable to the energy in the recording room.
- Step 4: Comparison of the IACC in the playback room and the IACC in the recording room. (arg min$(IACC_{rec}(n) - IACC_{play}(n))$)

The iterative process Step 2 to Step 4 is done for every frequency channel. After that, the final processing is made with the best suitable κ_n which minimizes the correlation difference between the recording and the playback room.

2.5. Synthesis

The synthesis stage is used as it was introduced by [8] and is shown in Figure 1. For the synthesis, a 4^{th}-order gammatone filterbank is used with a sampling frequency of 48 kHz. The filters have a bandwidth of 1 ERB (equivalent rectangular bandwidth) between 20 Hz and 20 kHz. This leads to 42 filter coefficients per channel. After processing the coefficients as described in Section 2.2, the coefficients β_n were applied in each gammatone band n as a real-valued gain factor for the direct sound. The same process will occur for the coefficients α_n for the dipole speakers. For the synthesis, the gammatone channels were phase aligned with a delay of 16 ms to avoid audible artefacts in the synthesis stage. The phase alignment is necessary to compensate the physiologically motivated delays of the auditory filters on the basilar membrane. [8] showed that a delay of 16 ms gives good results in this stage. After the phase alignment, the filtered impulse responses are summed across all filter channels P. An example for the direct sound $C(t)$:

$$C(t)_{opt} = \sum_{n=1}^{P} \beta_n \cdot \langle \gamma_n(C(t)) \rangle \tag{18}$$

Now, the direct sound $C(t)_{opt}$ can easily be played back in the playback room via the stereo loudspeakers. For a listening test it is possible to have a headphone reproduction such that a comparison can be made between the reference artificial head signal from the recording room with the artificial head signal of the playback room. This can be achieved by convolving the optimized direct sound with the BRIR of the front loudspeakers. For the headphone reproduction this procedure is done for the dipole loudspeakers, too.

3. RESULTS

3.1. Objective evaluation

In the following section the optimized parameters will be discussed. Figure 2 (top) shows the energy difference of the left artificial head ear between the recording room and the playback room for the simulated lecture room and different conditions. The red curve shows the error between the recording room and the playback room for the perceptual optimization. It can be seen that the fluctuations of the error is fairly small over a wide frequency range. For comparison two conventional recording methods are evaluated also. The first is a Room-in-Room (RinR) rendering which refers to a microphone placed at 2.6 m from the source in the recording room and for which the signal is rendered over two loudspeakers in the playback room. The multi-Channel (mCH) refers to a similar condition where the signals from the omni-directional microphones were rendered on two surround loudspeakers that were placed in the playback room (cf. Fig. 4). The comparison of the rendering methods RinR and mCH with the applied optimization shows that the fluctuations of both methods are greater than the applied optimization. In Figure 2 (middle) the interaural cross correlation is shown. By comparing the IACC of the recording room with the room-in-room (RinR) rendering method it can be seen that the IACC of the room-in-room method is higher over the whole frequency range. The multi-channel reproduction is much lower than the simple RinR-method, but it does not fit to the IACC of the recording room. The comparison of the optimized IACC with the IACC in the recording room shows that both curves fit quite well in most of the gammatone channels. In some channels the IACC cannot reach the desired correlation with simply the dipole signals. For a better adjustment of the correlation, a compensation with the front loudspeaker signals should be taken into account to achieve the reference IACC. In Figure 2 (bottom) the energy decay curve (edc) is illustrated for the recording room and the conditions Opt, RinR and mCH. The RinR-method shows that the edc has a higher descending slope and has a $T_{60,RinR} = 597$ ms. The multi-channel conditions shows a similar slope with a small offset like in the recording room and a $T_{60,mCH} = 695$ ms. The applied optimization shows that the the slope as well as the reverberation time $T_{60,Opt} = 706$ ms is comparable with the reverberation time in the recording room $T_{60,Rec} = 699$ ms.

Figure 3 (top) shows the energy difference of the left artificial head ear between the recording room and the playback room for the simulated church and different conditions. The error of the applied optimization is rather small over a wide frequency range. As can be seen the error between the recording room and the playback room for the conditions RinR and mCh is relatively high, which results from interference of room modes in both rooms. The difference among the RinR and mCh condition is fairly small because of the high front to back ratio of 10 dB, which leads to a similar IACC in Figure 3 (middle). The optimized IACC shows a good agreement with the IACC of the recording room. The energy decay curves illustrate similar properties. The $T_{60,Opt} = 3033$ ms in the playback room are in good agreement with the

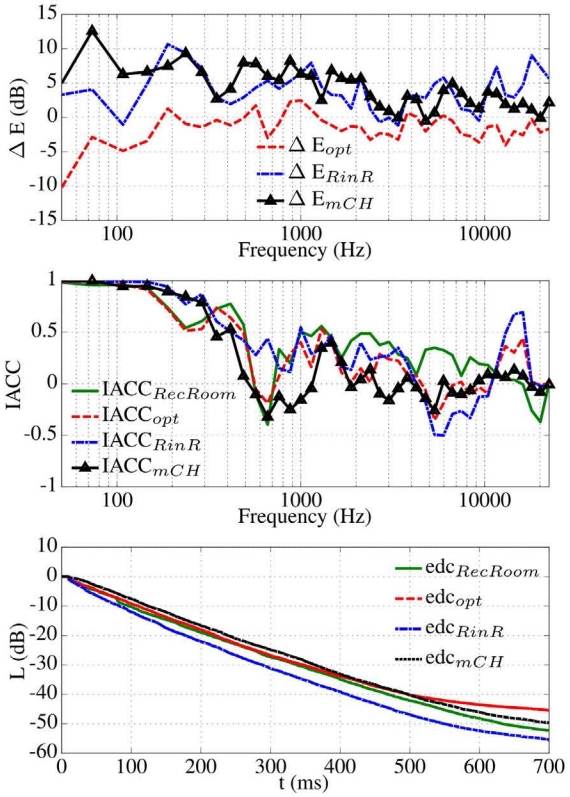


Figure 2: *Comparison of the parameters for the simulated lecture room. Top: Energy difference between the recording room and the playback room for the perceptual optimization (Opt, red), room-in-room (RinR, blue) and the multi-channel condition (mCh, black). Middle: Illustrated is the IACC of the recording room (green), the perceptual optimization (Opt, red), room-in-room (RinR, blue) and the multi channel condition (mCh, black). Bottom: Illustrated is the energy decay curve (edc) for the recording room (green), the perceptual optimization (Opt, red), room-in-room (RinR, blue) and the multi channel condition (mCh, black).*

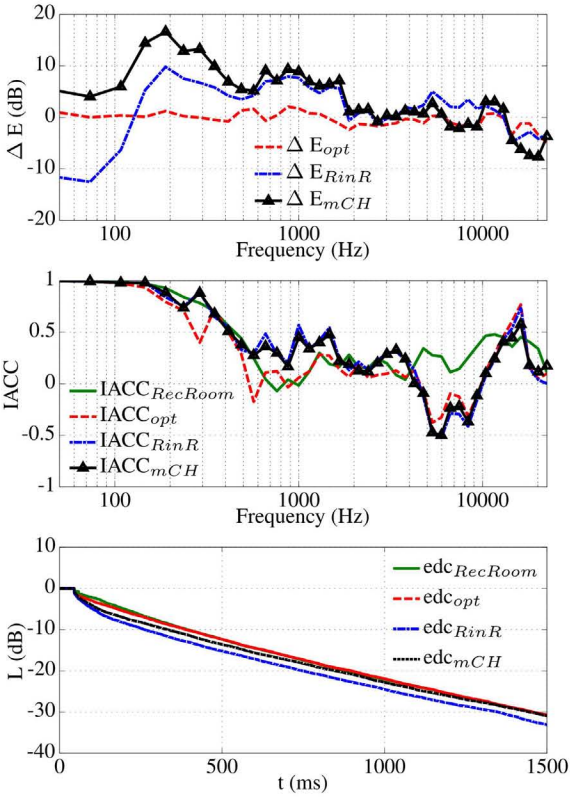


Figure 3: *Comparison of the parameters for the simulated church. Top: Energy difference between the recording room and the playback room for the perceptual optimization (Opt, red), room-in-room (RinR, blue) and the multi-channel condition (mCh, black). Middle: Illustrated is the IACC of the recording room (green), the perceptual optimization (Opt, red), room-in-room (RinR, blue) and the multi channel condition (mCh, black). Bottom: Illustrated is the energy decay curve (edc) for the recording room (green), the perceptual optimization (Opt, red), room-in-room (RinR, blue) and the multi channel condition (mCh, black).*

$T_{60,Rec} = 3040$ ms in the recording room. The multi-channel condition has a $T_{60,mCH} = 2768$ ms and the room-in-room reproduction a $T_{60,RinR} = 2921$ ms. The comparison of the T_{60} reverberation time at the artificial head between the recording room and the playback room with the applied optimization showed that it strongly depends on the separation time constant t_m which was introduced in section 2.1. It was found that for the simulated lecture room, a separation time constant of $t_m = 28$ ms gives a good approximation of the reverberation time of $T_{60,Opt} = 706$ ms which is 7 ms above the reverberation time of the recording room. For the simulated church an separation time constant of $t_m = 60$ ms was found which leads to a reverberation time in the playback room of $T_{60,Opt} = 3033$ ms which is 7 ms below the reverberation time of the recording room. The reproduced T_{60} are below the just noticeable difference of reverberation time, which is in the range of 20% to 30 % [10]. The optimal separation time constant t_m can be derived iteratively by varying t_m from small to bigger values. A small t_m means that only a small amount of the direct sound energy in the recording room is considered which leads to a small amount of direct sound in the playback room. Because the overall energy in the playback room should be comparable to the energy in the recording room, a larger amount of energy has to be rendered over the dipole loudspeaker. This leads to a higher reverberation time in the playback room compared to the T_{60} of the recording room. The optimal constant t_m is derived when the T_{60} in the playback room is comparable to the T_{60} in the recording room.

3.2. Subjective evaluation

In the following section the experimental setup of the listening experiment will be introduced. In this listening test, two recording rooms were simulated. The first room was a lecture room at the University of Oldenburg ($T_{60} = 650$ ms), the second recording room was the St.Marien Church in Oldenburg ($T_{60} = 3040$ ms). The playback room was the loudspeaker lab. ($T_{60} = 400$ ms) at the University of Oldenburg. The loudspeaker orientations used are shown in Figure 4 for the different test conditions. In the subjective evaluation a MUSHRA-Test was used to evaluate the different rendering methods over headphone relative to the reference condition (called ref) in the recording room. In addition, a conven-

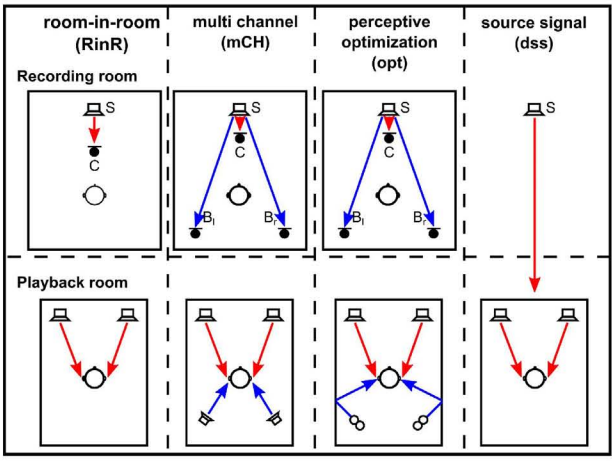


Figure 4: *The four conditions of the listening test. Room-in-room: The microphone signal $C(t)$ (red) is rendered over the front loudspeaker (red) in the playback room. Multi channel: The close microphone signal $C(t)$ (red) is rendered over the front loudspeaker (red) and the microphones $B(t)$ in the diffuse field (blue) are rendered over two rear-loudspeaker (blue). Opt: The optimized close microphone signal $C(t)$ (red) is rendered over the front loudspeaker (red) and the optimized microphones $B(t)$ in the diffuse field (blue) are rendered over the two dipole loudspeaker (blue). dss: This condition contains just the room-acoustical properties of the playback room. The dry source signal is rendered directly in the playback room over the front loudspeaker (red).*

tional room-in-room reproduction (called RinR) method was used. In this condition, the close cardioid microphone C in Figure 4 was recorded in a distance of 2.6 m to simulate a conventional stereo reproduction with a small amount of reverberation. This signal was rendered over the front loudspeaker. For the multi-channel reproduction (mCH) the close cardioid microphone signal C was recorded in a distance of 1.4 m. This signal was rendered over the front loudspeaker like in a 5.1 setup. The signals B_l, B_r was rendered as the two rear-speaker-signals of a 5.1 multi-channel setup. In the multi-channel condition no subwoofer and no center speaker was used. The applied optimization (Opt) was processed like it is described in Section 2 and then rendered in the playback room. As the anchor test-condition (anchor) the reference signal was low-pass filtered at 3.5 kHz as described in [ITU-R BS.1534-1]. The secondary anchor test-condition (called dss) is the dry instrument played back in the playback room. This condition was used to investigate the change in the perceived room-acoustics with respect to the other reproduction methods. To use the source signals of the recording room for the multi-channel reproduction, the energy-ratio of the front channels to the rear channels of four musical DVD's were analyzed. A front to back ratio of 10 dB was found. The close microphone signal of the front loudspeaker was recorded at the same distance which was used in the condition Opt. The dipole-speakers were replaced by two loudspeakers with a conventional directivity pattern (Genelec 6010A) and positioned as described in the [ITU-R BS.775].

3.3. Stimuli and subjects

Twelve different monaural recordings of musical instruments of five to ten seconds in duration were used. The instruments were dry music signals recordings without any room influences. The recordings used were as follows: a piece of Beethoven (recorded by [11]), a choir (recorded by [12]), female speech, a violin (one self recorded and one recorded by [12]), two guitars (chords and picking), clarinet, piano, saxophone, snare drum and a trumpet. All stimuli were presented at 67 dB-SPL. All stimuli were convolved with the room impulse response of the close microphone C and the microphone B in the recording room. These signals were then convolved with the specific binaural loudspeaker impulse response which was measured from the loudspeaker to the artificial head in the playback room (C with the BRIR of the front loudspeakers and B with the BRIR of the dipole loudspeakers). The same procedure was done for the listening conditions RinR and mCh. To have the possibility to compare the recording room with the playback room, the original source signal was convolved with the BRIR of the artificial head of the recording room as a reference signal. The listening test were performed by $N = 12$ normal hearing subjects, nine male and three female, with a mean age of 29 years. Five of twelve participants reported to have musical experience with playing an instrument. The rating was done from all subjects for all conditions and instruments in two sessions. The duration of one session was approximately 60 minutes. The task of the subjects was to rate in a blind test the difference of five processing algorithms (anchor (low-pass filtered at 3.5 kHz), Opt (perceptual optimized room-in-room reproduction), dss (dry signal in playback room), RinR (a conventional room-in-room reproduction) and a multi-channel reproduction) on a scale between 0 (large difference) and 100 (no difference) . Additionally, a hidden reference condition was included. All subjects completed a training phase where all stimulus manipulations were presented for a select number of test stimuli.

3.4. Subjective results

Figure 5 shows the results for the MUSHRA-Test for the lecture room (red) and the church (blue). Illustrated in Figure 5 is the mean over all subjects. The standard error is derived from the mean scores calculated over all subjects and thus shows the variations between the instruments. Examination of the data in Figure 5, it shows that our proposed method, Opt, was always rated with a smaller difference than the conventional room-in-room reproduction (RinR). This trend can be seen for the lecture room as well as for the church. The perceived difference could be caused by the stronger variations in energy (which cause an increase in coloration), a much higher IACC over all frequencies and a lower reverberation time which are illustrated in Figure 2. A comparison of the results of the condition Opt with the multi-channel condition (mCH) shows that for the lecture room the perceived difference is comparable. This can be seen in Figure 2, that the Energy Decay Curve of this condition is comparable to EDC of the recording room as well as the IACC. Comparing the conditions Opt and mCh of the church, it shows that the condition Opt was rated with no difference. The condition mCh shows that it was rated much lower as the RinR condition in the church. The reason why the condition RinR is rated much higher than the multi-channel (mCh) reproduction could be explained considering that the front to back ratio of 10 dB is too high. A reason for this is that the distance of the close microphone signal is closer at the sound source than the RinR condition. Therefore, the mCH condition has less reverberation in the close microphone signal, which cannot be compensated by the rear-speaker in the mCH condition. In the multi-channel condition, the ratio between the front and back channel signals control the direct to diffuse ratio of the rendered signal. The reproduced

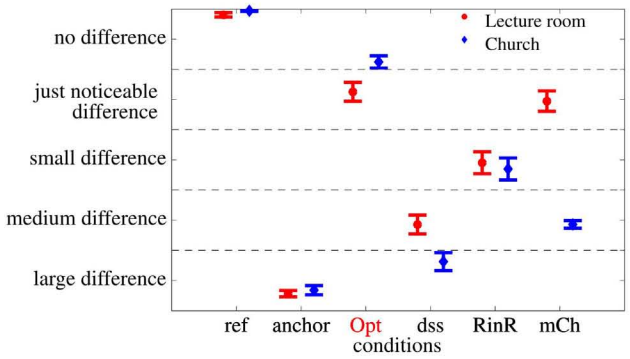


Figure 5: *Subjective measurement results for the headphone reproduction of the lecture room (red dots) and the church (blue diamonds) in a MUSHRA-Test. The symbols are the mean over 12 instruments and 12 subjects. The error bars are the standard error and indicates the variation over the 12 instruments. The x-axis are the different processing conditions, the y-axis indicates the difference on a scale between* 0 *(large difference) and* 100 *(no difference). Our proposed method (Opt) is marked in red on the x-axis.*

signals in the condition (mCH) are less reverberant than the condition RinR. The condition dss shows the dry source signal which was played directly in the playback room. This condition shows how large the perceived difference of the recording room is compared to the playback room. This condition should be rated much lower than the other conditions (with the exception of the 3.5 kHz anchor signal (anchor)) because only the room-acoustical properties of the playback room is included. In addition to the previous listening test, a live listening test over loudspeaker should be performed, to validate the previous results. This could be necessary to include the effects of head movements and the individual head-related-transfer-functions of the listener. Olive et al. ([13]) compared a live loudspeaker reproduction with a binaural reproduction over headphones in a subjective listening test. They showed that the scores between a live representation and a headphone representation have minor discrepancies which could result out of the removal of the visual biases and head movements. However, it can be seen that the standard errors are in a fairly small range and our proposed method works quite well over all stimuli used.

4. CONCLUSIONS

In this study a method was presented for rendering the room acoustics of a recording room in an echoic playback room. This method compensates for the reproduction conditions in the playback room. Rather than attempting to recreate the physical sound field, the proposed method optimizes the perceptual attributes IACC and the overall timbre in a playback room using a fairly small amount of loudspeakers. Because of the placement of an artificial head at the recording side it is possible to analyse the specific room dependent timbre and binaural cues and reproduce these on the playback side. For sound reproduction in the playback room, a conventional stereo loudspeaker setup is first used to reproduce the direct sound. With this setup, we can control the direction of arrival and the amount of energy in the playback room corresponding to the direct energy in the recording room. Furthermore the energy difference as well as the IACC can controlled with diffuse sound via a set of dipole loudspeakers. Because of the directivity pattern only the diffuse field is excited which implies that the dipoles are not perceived as a separate sound source provided that the head is in the sweet spot. For a better understanding in terms of head movements with this setup, a new listening test which covers various listening positions should be conducted. A comparison of the subjective ratings in Section 3.4 showed a higher preference for our proposed method with reference to the conventional room-in-room reproduction.

5. ACKNOWLEDGMENTS

This work was supported by the DFG Forschergruppe Individualisierte Hoerakustik (FOR-1732). The author thanks the two anonymous reviewers for the comment, that helped to improve the paper.

6. REFERENCES

[1] M.A. Gerzon, "Periphony: With-height sound reproduction," *J. Audio Eng. Soc.*, vol. 21, pp. 2–10, 1973.

[2] A.J. Berkhout, "A holographic approach to acoustic control," *J. Audio Eng. Soc.*, vol. 36, pp. 977–995, 1988.

[3] M.R. Schroeder, "Statistical parameters of the frequency response curves of large room," *J. Audio Eng. Soc.*, vol. 35, pp. 299–306, 1987.

[4] W. de Bruijn, A. Härmä, and S. van de Par, "On the use of directional loudspeakers to create a sound source close to the listener," in *124th AES Convention*, May 2008, Amsterdam, Netherlands.

[5] J. Breebaart, S. van de Par, A. Kohlrausch, and E. Schuijers, "Parametric coding of stereo audio," *Journal on Applied Signal Processing*, vol. 9, pp. 1305–1322, 2005.

[6] J.S. Bradley and G.A. Soulodre, "Objective measures of listener envelopment," *J. Acoust. Soc. Am.*, vol. 98, pp. 2590–2597, 1995.

[7] Brian C. J. Moore, *An Introduction to the Psychology of Hearing*, 6th edition edition, 2 January 2012.

[8] V. Hohmann, "Frequency analysis and synthesis using a gammatone filterbank," Tech. Rep., Acta Acustica united with Acustica , Vol. 88, pp. 433–442, 2002.

[9] J. Dattorro, "Constrained least squares fit of a filter bank to an arbitrary magnitude frequency response," 1991.

[10] Zihou Meng, Fengjie Zhao, and Mu He, "The just noticeable difference of noise length and reverberation perception," in *Communications and Information Technologies, ISCIT '06. International Symposium on,*, 2006, pp. 418 – 421.

[11] J. Pätynen, V. Pulkki, and T. Lokki, "Anechoic recording system for symphony orchestra," *Acta Acustica united with Acustica*, vol. 94, pp. 856–865, 2008.

[12] R. Freiheit, "Creating an anechoic choral recording," in *Proc. of the International Symposium on Room Acoustics*, 2010.

[13] Sean E. Olive and Peter L. Schuck, "The effects of loudspeaker placement on listener preference ratings," *J. Audio Eng. Soc.*, vol. 42, no. 9, pp. 651 – 669, 1994.

[14] O. Warusfel and N. Misdariis, "Directivity synthesis with a 3d array of loudspeakers application for stage performance," *Proceedings of the COST G-6 Conference on Digital Audio Effects*, vol. DAFX-01, pp. 1–5, 2001.

Proc. of the EAA Joint Symposium on Auralization and Ambisonics, Berlin, Germany, 3-5 April 2014

SCALE MODEL AURALIZATION FOR ART, SCIENCE, AND MUSIC: THE STUPAPHONIC EXPERIMENT

Brian F.G. Katz
Audio & Acoustics Group
LIMSI-CNRS
Orsay, France
brian.katz@limsi.fr

Markus Noisternig
Acoustic and Cognitive Spaces Group
UMR STMS, IRCAM-CNRS-UPMC
Paris, France
Markus.Noisternig@ircam.fr

Olivier Delarozière
Woodstacker
Champ-au-Beau, France
olivier.delaroziere@woodstacker.net

ABSTRACT

The use of acoustical scale models has been replaced for the most part by computational models and numerical simulations for room acoustic studies as well as artificial reverberation units. There remains however a number of acoustical phenomena which are difficult to address with computer simulations, such as coupled volumes, diffraction, and complex scattering, due to the computational complexity and/or calculation time necessary for addressing such acoustical wave phenomena on the scale of room acoustical problems, even small rooms. This paper presents a pilot study of a rather unique artistic architectural structure consisting of a self-supporting construction composed of small stacked linear elements. Acoustically, the structure combines modal behavior, concave forms, and very regular scattering patterns. An example scale model has been constructed and studied in order to separate different construction features and their associated acoustics effects. In an attempt to explore the interest of the specific acoustic for musical performance, a computational platform was created to utilize the scale model as a physical convolution reverberation unit for musical performance.

1. INTRODUCTION

With the advent of recording, and dry recording studios, there have been many efforts developed for the reintroduction of reverberation into studio recorded music. Some of the first technologies developed were the use of "echo chambers", wherein the dry audio captured with the microphone was diffused in a reverberant environment over loudspeakers, and then recaptured with microphones. This physical-based artificial reverberation was quite popular, with examples existing in such famous institutions as Abbey Road Studios where the echo chamber was constructed in 1931[a,b]. Echo chambers are, however, space demanding, difficult to transport, and not extremely adjustable. With improvements in electronics, other physical-electronic reverberation systems have been developed such a plate reverberators and spring reverberators.

[a] http://en.wikipedia.org/wiki/Echo_chamber, last viewed 2013-11-30

[b] http://audiogeekzine.com/2011/02/the-history-of-echo-echo-chambers-chambers/, last viewed 2013-11-30

With improvements in computer processing power, purely electronic reverberation became possible, such as using feedback delay network (FDN) for reverberation processing [1, 2, 3]. These reverberators could be easily adjusted, for example using perceptual descriptors relying on a simplified model of the time-frequency energy distribution of parametric FDN [4]. Such reverberators are however limited, lacking certain realism and ability to represent unique architectural elements. Additional increases in computational power allowed for the use of convolution reverberators, using complex impulse responses, either measured or calculated based on geometrical models such as ray tracing [5, 6], beam tracing [7, 8, 9], or radiosity [10, 11]. Convolution reverberators capture the fingerprint of a given space, but require preparations for the acquisition of such IRs and allows little flexibility with regards to modifying the room at time of use, though perceptual control of convolution based room simulators is a subject of current study [12].

Scale models, to date, have been used as off-line convolution reverberators to study architectural spaces [13, 14], but never in a performance setting. The current study envisages the possibility of using scale models in the same way as the old "echo chambers" of the early and mid-twentieth century, while allowing for the creation of complex and unique acoustic spaces rather than just simple reverberators. Real-time use of one, or several, models and the ability to dynamically alter source, receiver, and even room positions and configurations as desired offers a new form of reverberation and musical expression.

In parallel, the development and exploitation of real-time scale model convolution offers a number of interesting scientific aspects. To begin with, there is the basic signal processing challenge of achieving such a system. The applications of real-time physical-based convolution in scale models, in contrast to off-line convolution with measured impulse responses of the scale model, offers the ability to study room excitation by dynamic sources, such as moving or rotating, with perceptual studies. of specific interest are perceptual studies concerning musician/room interactions, which require real-time processing of generated music in coordination with source dynamics. The effect of dynamic architectures can also be examined, such as movable panels, or dynamic listener placement or movement during a performance.

2. ARTISTIC CONTEXT

What child has not dreamed of being able to experience, as a Lilliputian, a world in miniature: doll houses, electric trains, miniature circus . . . In the "Stupaphonic" [1] project that childhood dream will become a reality for musicians: they will be able to play to their audience in a space in miniature.

This particular space is based on a special type of structure, which is at the core of the architectural project *Woodstacker* [15]. This architectonical type of structure is a solution to the geometrical problem of how to cover a large area by stacking small pieces of wood without the use of glue or nails. The result is a bottle shaped three-dimensional rose window (see Figures 1 and 4). This new building system, based on "stacked laminated" timber structures, can evoke references to *pagodas* whose construction also consists of wooden stacked elements. This stacked architecture, like *chorten* of Tibet, belongs to a family of stacked structures which are derived from a Buddhist mound-like structure called *stupa*. Stupas originated as pre-Buddhist earthen burial mounds, like *tumuli* in Europa. Thus was born the idea of linking our new project to these ancient and universal architecture.

The stupa is used by Buddhists as a place of meditation. In the original pre-Buddhist burial mounts ascetics were buried in a seated position. The American anthropologist J. Jaynes [16] proposes that they were buried in this position so they can continue to speak to living people. Hearing voices from beyond the grave suggests some acoustic illusions which are also a part of our device. In the Woodstacker system, the special geometrical pattern of the lamellas not only focuses the sound [17] but also functions as frequency filters producing a particular, almost metallic, sound. This strange acoustic effect is a second reason to link our project to the *stupa* as a kind of container for "voices from beyond the grave", a "voice granary" ("grenier à voix" in French) to cite the french writer Pascal Quignard [18].

The larger structures we have currently built can accommodate about 30 people for sound experiments (see photograph in Figure 1). This size limitation is a compromise between the funding for artistic experiments and the cost of such a construction. With the project's evolution we desired a means to quickly experiment with different architectural structures in a flexible way. The use of physical scale models which are powerful tools for architects, carpenters, and acousticians [19] offered a solution to exceed the current constraints. Thus, we found a way to invert the acoustical environment, like turning a glove inside-out, and give the musician and audience located outside of the building the same acoustical experience that they could have inside the structure. Using the acoustic scale effect we are able to drastically reduce the size of our installation and increase the number of structures performers can play with and turn "stackscapes" (see Figure 2) into interactive soundscapes.

Figure 1: *Photo of live performance at StackCamp 2013 featuring Didier Petit (cello) and Emre Gultekin (saz). Champ-au-Beau, France.*

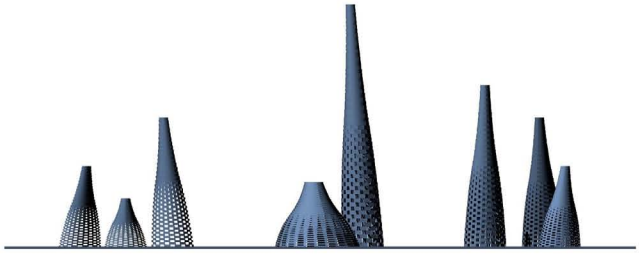


Figure 2: *Blue Stackscape, ©2006 O. Delarozière.*

3. SYSTEM OVERVIEW

To achieve the artistic, acoustic, and audio scheme imagined, a basic scenario and system architecture was envisioned. One can imagine a performance area, where the performer is equipped with one or several microphones. The space is open and large. Near the musician is one or several acoustic scale models, equipped with ultrasonic speakers and microphones. Around the musician is the audience. The sound produced by the musician is captured, transformed to the scale of the model where it is played and recaptured, then transformed to the full scale of the musician's performance, where it is played live to the audience over an electro-acoustic array of speakers either on-stage or around the audience.

3.1. Signal Processing

The signal processing chain is depicted in Figure 3. First, the input audio signal is up-sampled to the ultrasound sampling rate, which is determined by the scaling factor of the model. Second, the up-sampled input signal is transposed by the scaling factor preserving the harmonic structure of the signal. Two implementations have been tested: a) off-line transposition that allows for the time-stretching of the signal; b) a real-time implementation thereof, which compensates for the time-scaling effect using

[1] Stupaphonic: from stupa (from Sanskrit: m., स्तूप, *stūpa*, literally meaning "heap" [a]) and phonic (from Ancient Greek *φωνή*, *phōnē*, meaning "voice" or "sound" [b])

[a] http://en.wikipedia.org/wiki/Stupa, last viewed 2013-11-30

[b] http://en.wikipedia.org/wiki/Phonetic, last viewed 2013-11-30

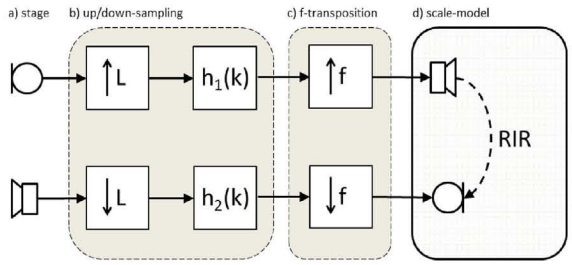


Figure 3: *Signal processing flow chart: (a) instrument signal capture, transformed audio signal playback; (b) up/down-sampling with anti-aliasing filters; (c) frequency transposition, and (d) the Stupaphonic scale model.*

a phase vocoder and thus preserves the continuity of the audio stream.

The off-line version study was carried out using MATLAB®. Audio samples were processed individually, with no audio streaming functionality. The basic approach consisted of taking an audio extract, resampling the audio, modifying the sample rate in order to apply the scale factor, play/recording the sample in the model, and then retransforming the recorded physically-convolved signal to full-scale for audio playback. A code sample for the described process is provided below:

```
fs = 44100;
[y] = wavrecord(audiolength,fs);
fs_max = 192000; % sample rate in scale model audio chain
scale_desired = 12;
[y_resamp] = resample(y,fs_max/scale_desired,fs);
fs_resamp = fs_max/scale_desired;
[y_convolved]=wavplayrecord(y_resamp,fs_max);
wavplay(y_convolved,fs_resamp);
```

In this example, with a maximum sample rate of 192 kHz on the audio system and a scale factor of 12, the recorded audio track is resampled to 16 kHz (bandlimited to 8 kHz). This resampled track is then played back and recorded in the scale model at an actual sample rate of 192 kHz. The recorded physically-convolved signal is then played back at an actual sample rate of 16 kHz, or resampled to the sample rate of the audio device. The simple redefinition of the sample rate for the audio buffer performs the application of the scale factor, while the resampling assures correct anti-aliasing filters.

While currently tested in single buffer full convolution, future studies will evaluate the possibility of applying the concept of overlap-add convolution [20] to the concept of this physical-based convolution in order to allow for real-time operation on audio streams.

The real-time version study was conceived of as an alternate approach to the above approach employing resampling and transposition to apply the scaling factor through the use of a high quality phase vocoder architecture. Phase vocoder techniques are typically based on a sinusoidal signal model. The digital audio sampling rate conversion employed band-limited interpolation (see e.g. [21, 22]) that can be efficiently implemented with sinc-function look-up tables. In [23] it was shown that parametrized phase vocoders can also be applied to non-sinusoidal signals. However, initial tests showed that the sinusoidal signal model limits the use of phase vocoders for real-time scale-model processing for large scale factors. Modified algorithms are the subject of continuing investigations.

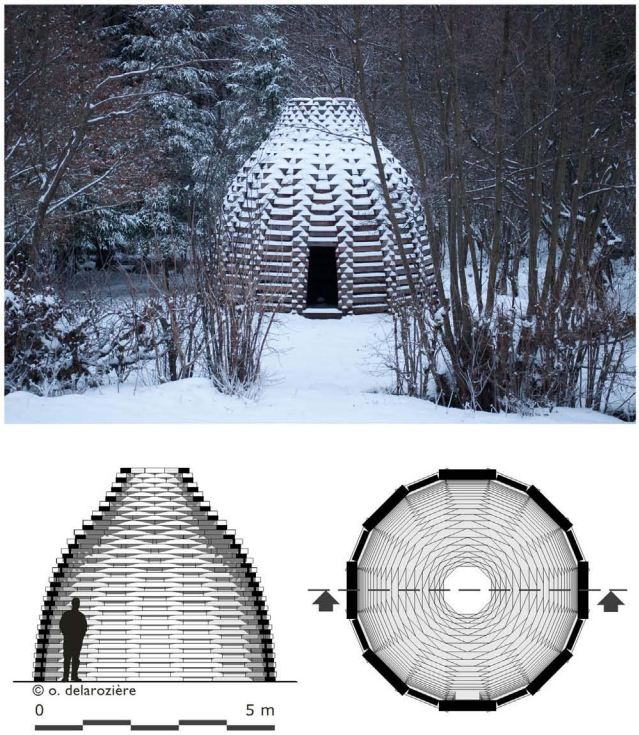


Figure 4: *Woodstacker stacked lamella timber cupola. Champ-au-Beau, France, 2010. (upper) A winter outside view. (lower) Section and Reflected ceiling plan.*

3.2. Scale Model

This preliminary study has been carried out using a single structure as a test case. The scale 1:1 structure was built in 2008 for a Land Art Exhibition in the highland of Auvergne, France (see Figure 4). The original building comprised 366 pieces of Douglas pine wood. It was 5 m in diameter, 4.5 m high, and weighed 6 tons. This work called "Vox Granarium" [24] was dedicated to famous ancient fiddlers from this area. This installation was dismantled and moved to Morvan where it was rebuilt in 2010. The Stupaphonic model is a 1:12 scale model of "Vox Granarium", 425 mm in diameter and 322 mm high. This scale was chosen due to material availability and because it is a traditional doll house scale. Serendipitously, 1:12 is also the Lilliputian people's scale in Gulliver's Travels[2]. The model was constructed from oak wood, whose lengths and widths were hand cut with no automatic process used for assembly. Unlike the full-scale construction, glue was used to fix the lamellas, and the model was assembled in three parts for ease of transportation and manipulation purpose (see Figure 5).

Due to the very long time needed to build this model by hand, subsequent models will probably be built using rapid prototyping techniques such as laser cutting. This will allow to quickly experiment with a large variety of structural shapes and configurations. We are also planing to use other materials such as metal or concrete for special acoustics effects.

[2] ". . . having taken the height of my Body by the help of a *Quadrant*, and finding it to exceed theirs in the Proportion of twelve to one . . ." [25, p. 64]

Figure 5: *Scale model (1:12) of "Vox Granarium", highlighting the tetrahedral acoustic source and 3 modular elements.*

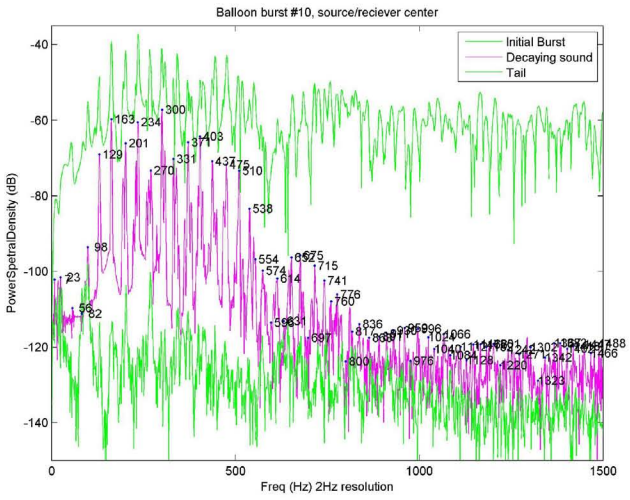


Figure 6: *Spectral analysis of impulse response (balloon burst excitation) of the full (1:1) scale "Vox Granariium", recorded with a sample rate of 96 kHz. Temporal analysis windows corresponding to direct/initial* $(0 - 10$ *msec), early decay* $(10 - 30$ *msec), and late tail* $(30 - \infty$ *msec).*

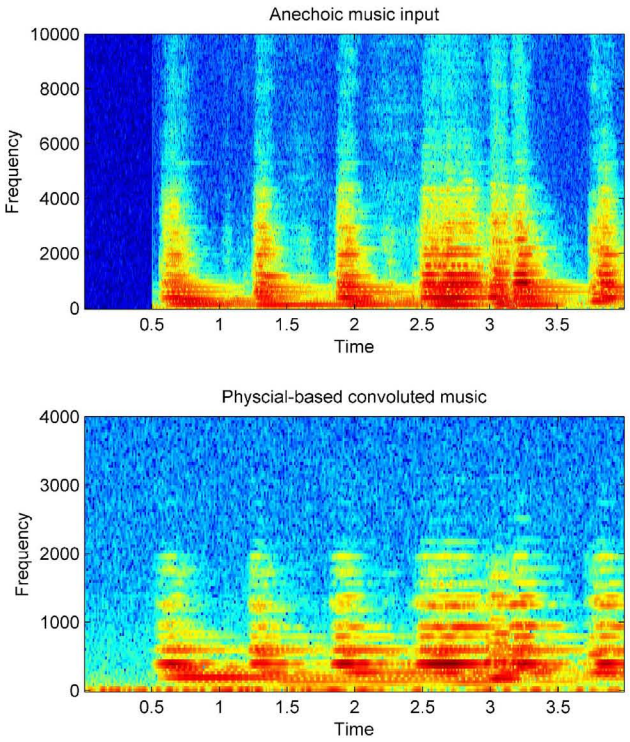


Figure 7: *Spectrogram of anechoic music extract (upper) and physical-based convolved music (lower) manually time-aligned.*

4. PRELIMINARY TEST

The diffraction acoustic effect discussed in Section 2 can be observed as a series of high-Q total resonances. These resonances can be observed by comparing the spectral response (magnitude of the FFT) at different moments in the impulse response. Figure 6 shows the spectral response at three moments in the impulse response (balloon burst excitation signal employed for room acoustics measurements [26]) of the full (1:1) scale "Vox Granariium" (see Figure 4). Resonant peaks are identified in the response for identification. There is clearly a region of resonance density over the frequency range 100–600 Hz, continuing still to $\approx$ 800 Hz.

The audio system employed for use in the scale model consisted of DPA 4060 microphones and a custom 3-speaker tetrahedron (see Figure 5) driven by a Samson Servo amplifier, connected to a RME Fireface 400 audio interface. While somewhat unconventional in traditional scale model research, this selection of pro-audio equipment has been used in previous studies [27, 28] and has been shown to provide improved signal-to-noise ratio when compared to more traditional laboratory scale model measurement architectures. The current hardware exhibits a frequency roll-off at $\approx$ 50 kHz. This of course imposes a low-pass frequency limitation for the physical-based convolution. With a scale factor of 12, the upper frequency limit due to this roll off is on the order of 4.2 kHz, rather than the 8 kHz permissible due to sampling theory. While suitable for the majority of studies in room acoustics with scale models, the musical implications of this limit can not be ignored. This limit can be raised by improving the upper frequency limit of the audio chain or selecting a lower scale factor.

An example result of the processing chain can be seen in Figure 7, which shows the spectrogram of a dry music extract before and after the physical-based convolution processing. The test music except was a dry multichannel recording of a Schubert trio (D.929, op.100), by [29] and publicly available [a].

The acoustic timbre of the convolved signal using the scale model greatly resembled that of the musical experience heard within the full scale installation. Even thought the processing steps apply a low-pass filter effect, due predominantly to transducer and amplifier performance limitations above 50 kHz, the frequency range where the resonance characteristics of the structure are apparent are still well within the operating frequency of the current signal processing chain for the 1:12 scale.

5. CONCLUSIONS

This paper has presented the foundations of the "Stupaphonic experiment", an artistic and scientific project which aims to use scale models as physical-based convolution reverberators. The architectural structure at the center of the project offers specific timbral qualities which are maintained in the initial tests, despite the frequency limitations of the scale transformations and associated signal processing chain.

The current example operates in an off-line, or time-deferred situation. While streaming is currently still being investigated, the current implementation could still be used in a performance setting

[a] http://c4dm.eecs.qmul.ac.uk/rdr/handle/123456789/27

in a live looping context, where the musician could send different audio samples to different architectures at a unique or different scale factors, effectively changing the size of the "echo chamber".

The development of the real-time processing stage, currently a subject of study, will allow for exploitation of the proposed physical-based convolution for studies in room acoustics, specifically those involving dynamic source, listener, or architectural elements, as well as dynamic performer/room interactions.

One artistic performance aspect specific to this system is the potential for cross-scale cross-talk. If the scale models are open to some degree, then the up-scaled audio will be heard by some of the audience. At the same time, full scale sounds, such as other elements of the performance or noise from the audience, can also be captured in the scale model, and subsequently down-scaled and played over the reproduction array. According to the Lilliputian scale factor of 1:12, one can imagine the majority of these sounds will be shifted to the lower end of the audible range, or into the subsonic region. However, a high pitched scream with a center frequency in the 5 kHz third-octave band for example would be clearly audible when transposed to the 400 Hz third-octave band, albeit also time stretched to 12 times its original duration. Investigations of these effects, and their possible artistic use, remain the subject of further studies.

6. ACKNOWLEDGMENTS

This study was funded in part by an Action Initiative grant at the LIMSI-CNRS.

7. REFERENCES

[1] J-M. Jot and A. Chaigne, "Digital delay networks for designing artificial reverberators," in *Proc. 90th AES Convention*, Paris, France, Feb. 1991.

[2] G. Garcia, "Optimal Filter Partition for Efficient Convolution with Short Input/Output Delay," in *Proc. 113th AES Convention*, Oct. 2002.

[3] M. Noisternig, A. Sontacchi, T. Musil, and R. Höldrich, "A 3D Ambisonic based Binaural Sound Reproduction System," in *Proc. 24th AES Int. Conf.*, Banff, Canada, June 2003.

[4] J-M. Jot, "Real-time spatial processing of sounds for music, multimedia and interactive human-computer interfaces," *Multimedia Systems*, vol. 7, no. 1, pp. 55–69, 1999.

[5] A. Krokstad, S. Strom, and S. Sorsdal, "Calculating the acoustical room response by the use of a ray tracing technique," *J. Sound Vib.*, vol. 8, no. 1, pp. 118–125, 1968.

[6] M. R. Schroeder, "Digital Simulation of Sound Transmission in Reverberant Spaces," *J. Acoust. Soc. Am.*, vol. 47, no. 2, pp. 424–431, 1970.

[7] T. A. Funkhouser, I. Carlbom, G. Elko, G. Pingali, M. Sondhi, and J. West, "A beam tracing approach to acoustic modeling for interactive virtual environments," *Proc. ACM Comp. Graphics (SIGGRAPH'98)*, pp. 21–32, July 1998.

[8] S. Laine, S. Siltanen, T. Lokki, and L. Savioja, "Accelerated beam tracing algorithm," *Applied Acoustics*, vol. 70, no. 1, pp. 172–181, 2009.

[9] M. Noisternig, B. F.G. Katz, S. Siltanen, and L. Savioja, "Framework for real-time auralization in architectural acoustics," *Acta Acoust. united with Acust.*, vol. 94, pp. 1000 – 1015, 2008.

[10] C. Malcurt, *Simulations informatiques pour prédire les critères de qualification acoustique des salles. Comparaison des valeurs mesurées et calculées dans une salle à acoustique variable*, Ph.D. thesis, Laboratoire Acoustique Métrologie Instrumentation, Toulouse, France, July 1986.

[11] G. I. Koutsouris, J. Brunskog, C-H. Jeong, and F. Jacobsen, "Combination of acoustical radiosity and the image source method," *J. Acoust. Soc. Am.*, vol. 133, no. 6, pp. 3963–3974, 2013.

[12] T. Carpentier, T. Szpruch, M. Noisternig, and O. Warusfel, "Parametric control of convolution based room simulators," in *Proc. Int. Symp. on Room Acoust. (ISRA)*, Toronto, Canada, June 2013.

[13] Jean-Dominique Polack, Xavier Meynial, and Vincent Grillon, "Auralization in scale models: Processing of impulse response," *J. Audio Eng. Soc*, vol. 41, no. 11, pp. 939–945, 1993.

[14] Vincent Grillon, Xavier Meynial, and Jean-Dominique Polack, "Auralization in small-scale models: Extending the frequency bandwidth," in *Audio Engineering Society Convention 98*, Feb 1995.

[15] O. Delarozière and U. Gleeson, "Woodstacker," in *Architectures autrement : Habiter le monde*, M. Culot and A-M. Pirlot, Eds., pp. 46–51. AAM, Brussels, 2005.

[16] J. Jaynes, *La naissance de la conscience dans l'effondrement de l'esprit*, Presses Universitaires de France, 1994.

[17] B. Katz, O. Delarozière, and P. Luizard, "A ceiling case study inspired by an historical scale model.," in *Proc. 8th Int. Conf. on Auditorium Acoust., Institute of Acoustics*, Dublin, May 2011, vol. 33, pp. 314–321.

[18] P. Quignard and C. Lapeyre-Desmaison, *Pascal Quignard le solitaire : Pascal Quignard, rencontre avec Chantal Lapeyre-Desmaison*, Les Flohic éditions, 2006.

[19] O. Delarozière, "Camera tectonica : Hypothèses pour un facsimilé d'architecture," in *Utopia Instrumentalis : Facsimilés au musée - Musée de la Musique*, Cité de la musique, Paris, Nov. 2010, pp. 46–56.

[20] A. V. Oppenheim and R. W. Schafer, *Digital signal processing*, Prentice-Hall, Englewood Cliffs, N.J., 1975, ISBN 0-13-214635-5.

[21] R. W. Schafer and L. R. Rabiner, "A digital signal processing approach to interpolation," *Proceedings of the IEEE*, vol. 61, no. 6, pp. 692–702, 1973.

[22] J. O. Smith, III and P. Gossett, "A flexible sampling-rate conversion method," in *Proc. IEEE Int. Conf. Acoust., Speech and Sig. Proc. (ICASSP)*, 1984, pp. 112–115.

[23] W.-H. Liao, A. Roebel, and A. W. Y. Su, "On stretching gaussian noises with the phase vocoder," in *Proc. of the 15 Int. Conference on Digital Audio Effects (DAFx-12*, Sept. 2012.

[24] O. Delarozière, "Vox Granarium," in *Horizons - Rencontres "Arts Nature"*, pp. 12–13. Office de Tourisme du Sancy, July 2008.

[25] J. Swift, *Part 1. A Voyage to Lilliput*, vol. 1 of *Travels Into Several Remote Nations of the World*, chapter III, pp. 47–64, Printed for *Benj. Motte*, at the Middle *Temple-Gate* in *Fleet-street*, 1726.

[26] J. Pätynen, B.F.G. Katz, and T. Lokki, "Investigations on the balloon as an impulse source," *J. Acoust. Soc. Am.*, vol. 129(1), pp. EL27–EL33, 2011.

[27] Paul Luizard, *Les volumes couplés : comportement, conception, et perception dans un contexte de salle de spectacle*, Ph.D. thesis, Université Pierre et Marie Curie, Paris, France, 2013.

[28] P. Luizard, M. Otani, J. Botts, L. Savioja, and Brian F. Katz, "Comparison of sound field measurements and predictions in coupled volumes between numerical methods and scale model measurements," in *Proc. Meetings on Acoustics*, Montreal, June 2013, vol. 19, p. (9 pages).

[29] Joachim Fritsch, "High quality musical audio source separation," M.S. thesis, UPMC / IRCAM / Telecom ParisTech, 2012.

A COMPARISON OF LATE LATERAL ENERGY (GLL) AND LATERAL ENERGY FRACTION (LF) MEASUREMENTS USING A SPHERICAL MICROPHONE ARRAY AND CONVENTIONAL METHODS

David A. Dick and Michelle C. Vigeant

Graduate Program in Acoustics,
The Pennsylvania State University
University Park, PA, United States of America
dad325@psu.edu

ABSTRACT

Late Lateral Energy Level (GLL) and Lateral Energy Fraction (LF) are two room acoustics measures that have both been shown to correlate with certain aspects of the spatial impression of a listening space. In order to obtain these quantities, the lateral energy must be measured, which is typically carried out using microphones with a figure-of-eight (figure-8) polar pickup pattern. However, most commercially available figure-8 microphones are intended for use in audio recording applications, and are not laboratory-grade or designed for room acoustics impulse response (IR) measurements. Such microphones may suffer from non-ideal frequency response and/or directivity patterns. This study compares measurements that were taken in a 2500 seat auditorium using an omni-directional and studio-grade figure-8 microphone pair versus the omni-directional and dipole components extracted from a 32 element spherical microphone array. The results show that the two measurement methods agree in the 500 Hz and higher octave bands, but differ at low frequencies due to differences in the directivity patterns. The difference of the LF average from 125 Hz to 1 kHz for the two methods was between 0.59 and 1.81 just noticeable differences (JNDs) at the six receiver locations. The difference of the GLL average from 125 Hz to 1 kHz for the two methods was between 0.02 and 0.48 JNDs (applying the JND for strength of 1 dB). It was also found that repeatability error was present at one of the six receiver locations for the LF measurements, but was very small for the GLL measurements.

1. INTRODUCTION

To quantify certain spatial parameters in rooms, a measurement of the lateral sound field is required, which is typically acquired using microphones with a figure-8 polar pickup pattern. The focus of this study was to compare the measurement of spatial measures by obtaining the figure-of-eight (figure-8) room impulse response (IR) in two different ways: using a microphone with a figure-8 directivity pattern, and by using a spherical microphone array and obtaining the figure-8 pattern by beamforming the dipole response.

It should be noted that while the spherical microphone array used in this study is capable of a third order spherical harmonic expansion, only the first order dipole is required for these measurements. While the full capability of the microphone array isn't needed for this work, this study serves as a verification that spherical microphone arrays, which are now being used for advanced analysis of room IRs, can also be used to measure the metrics commonly used in the architectural acoustics community which are outlined in the room acoustics measurement standard, ISO 3382 [12].

1.1. Spatial Measures

There is ongoing research to find objective metrics that correlate with the subjective perceptions of the quality of concert hall acoustics. Spatial impression is one characteristic that has been shown to be related to overall quality [1]. This concept was explored in terms of early lateral reflections by Barron [2], and low frequencies were also found to be an important component of spatial impression [3]-[5]. Further research proposed that spatial impression should be formally divided into two distinct components [5], with the particular details established by Bradley and Soulodre who defined: the apparent source width (ASW) as being associated with the early lateral reflections, and listener envelopment (LEV), which is related to late lateral reflections [6].

A number of objective measures have been proposed to predict both ASW and LEV, but only two were the focus of this study. Lateral energy fraction (LF) is a commonly used parameter to predict the perception of ASW, which is the ratio of early lateral energy to total early energy [2]:

$$LF = \frac{\int_{5\ ms}^{80\ ms} p_f^2(t)dt}{\int_0^{80\ ms} p_o^2(t)dt} \quad (1)$$

where $p_f(t)$ is the room IR measured with a figure-8 microphone, and $p_o(t)$ is the room IR measured with an omnidirectional microphone. Late lateral energy level (GLL) has been used to predict LEV, which is the ratio of the late lateral energy to the normalized source energy [6]:

$$GLL = 10log\left[\frac{\int_{80ms}^{\infty} p_f^2(t)dt}{\int_0^{\infty} p_a^2(t)dt}\right] \text{ [dB]} \quad (2)$$

where $p_f(t)$ is the room IR measured with a figure-8 microphone, and $p_a(t)$ is the IR of the sound source normalized at a distance of 10 meters away in a free field.

1.2. Measurement Uncertainty of Spatial Measures

Only a handful of studies have been published that evaluate measurement uncertainty of spatial measures and the results from the majority of the studies show a high degree of uncertainty [7] - [10]. One of the earliest studies showed that the standard deviation

across the results from four measurement teams for LF measurements was up to 0.20 at 1 kHz which each used a different figure-8 microphone [7]. A second study compared the results of LF and GLL measurements from a figure-8 and omni-directional microphone pair and a custom intensity probe, and reported significant differences in the results likely due to variations in the microphone directivity patterns [8]. The first phase of the third room acoustics simulation programs round robin study was to collect measurement data on the space that was to be modeled. The results for the typical parameters, e.g. reverberation time (T30), early decay time (EDT), etc. were very similar across the four measurement teams [9]. However, significant differences in LF were found, which were on the order of 3 just noticeable differences (JNDs), where the JND for LF is 0.05. Some follow-up measurements using three figure-8 microphones of the same make and model (Neumann KM86) revealed significant differences in measurements taken with the microphones at different orientations. One possible source of this measurement error was hypothesized to be due to changes in the microphone sensitivity of each diaphragm due to aging.

A more recent case study was conducted to further evaluate the measurement uncertainty of spatial measures in terms of microphone orientation, spacing between the microphone pair, and microphone type [10]. A total of five different makes and models of figure-8 microphones were evaluated by taking measurements in a small lecture hall with about 100 seats. The average differences due to microphone spacing, which varied between 64 to 152 mm, and microphone orientation, were found to be relatively small for GLL, which were on the order of 0.2 dB, but were slightly higher for LF, on the order of 3 JNDs. On the other hand, the effect of microphone type was more significant for GLL, with variations on the order of 1.5 dB, and similar variation of about 3 JNDs for LF.

1.3. Microphone Limitations

For this experiment, three different microphones were used to measure room IRs: a Brüel & Kjær (B&K) Type 4192 omni-directional microphone, a Sennheiser MKH 30 Figure-8 microphone, and an mh Acoustics em32 Eigenmike® spherical microphone array. The Sennheiser microphone and the Eigenmike both have their own distinct disadvantages in measuring the lateral energy component of the IR. The Sennheiser microphone is not a laboratory-grade instrument, and the frequency response is not flat broad-band. In addition, the linearity of the microphone is not known, and was not measured as a part of this study. The Eigenmike has a high frequency limit of approximately 8 kHz due to spatial aliasing. Below approximately 150 Hz, the Eigenmike begins to veer away from an ideal dipole shape due to white noise gain constraints. Below this frequency, the null shifts in angle, and the main lobes are no longer symmetric [11].

2. MEASUREMENT PROCEDURE

2.1. Measurement Equipment

A B&K Type 4292-L OmniPower Sound Source dodecahedron loudspeaker was used for the source, driven with a Crown K2 amplifier. An RME Babyface was used for the audio interface with the computer for the B&K and Sennheiser pair. For the Eigenmike configuration, the Eigenmike Interface Box (EMIB) was used as the audio interface, and the RME Babyface was used as a D/A converter to send the output signals from the Eigenmike Interface Box to the amplifier. EASERA room acoustics software running on a MacBook Pro was used to measure the IRs. The EMIB connects to the MacBook via FireWire interface to send the 32 channels of data from the Eigenmike to the computer.

A custom microphone stand was built that could be used for both microphone configurations, the Eigenmike, and the omni and figure-8 pair (see Figure 1). The omni and figure-8 microphones were placed 7.6 cm apart from each other. This spacing was used to allow for the microphones to be adequately far enough apart so as to minimize the effects on the other microphone [10], but close enough so that they were measuring approximately the same point in space.

During the measurements, the base of the microphone stand was positioned in front of the seat, and the adjustable arm was used to place the microphone in the location of a listener's head. The microphone stand is adjustable in each dimension separately to allow for accurate and precise positioning of the microphone. For these measurements, the center of each microphone array (either the center of the Eigenmike array or the center of the two discrete microphones) was placed in the halfway across the width of the chair, 20 cm from the seat back, and 70 cm above the seat bottom.

Figure 1: *The two microphone arrays in the custom microphone stand.*

2.2. Anechoic Chamber Measurements

In order to calculate GLL, the free field sound pressure level of the sound source at 10 meters must be obtained to use in the denominator in equation (2), which is typically done using an anechoic chamber. An IR measurement was taken every 12.5 degrees around the dodecahedron loudspeaker at a distance of 3.55 meters away according to ISO 3382 [10] using a B&K 4191 free field microphone. The stimulus was a swept sine signal with a pink-weighted spectrum, which was played at the same level that was used for the IR measurements in the hall. The resulting 29 measurements were energy-averaged to account for the directivity of the source, and normalized to a distance of 10 meters away.

Since standard calibrators are not available for either the Eigenmike or the Sennheiser MKH 30, the microphones had to be calibrated using a loudspeaker playing a calibration tone in an anechoic chamber. A 1 kHz tone was played over the dodecahedron loudspeaker, and the sound level was measured using a calibrated sound level analyzer, B&K type 2250. These levels were entered into the measurement program, which used these quantities to calculate the microphone sensitivities. The Eigenmike includes a PC application which controls programmable gain amplifiers to correct for magnitude differences between the individual microphone capsules.

To verify the directivity of the Sennheiser and Eigenmike, the frequency response of each microphone was measured in an anechoic chamber as a function of angle in the horizontal plane. The microphone was placed on a turntable two meters away from a stationary loudspeaker, and the frequency response function was measured with each microphone rotated every three degrees.

2.3. Room Impulse Response Measurements

IR measurements were taken in the Eisenhower Auditorium located on The Pennsylvania State University campus in University Park, PA in the United States. The source was placed in the center of the stage for all measurements. Six receiver locations were chosen in the hall: two on the main floor (R1 and R2), two on the grand tier level (R3 and R4), and two on the balcony level (R5 and R6), as shown in Figure 2.

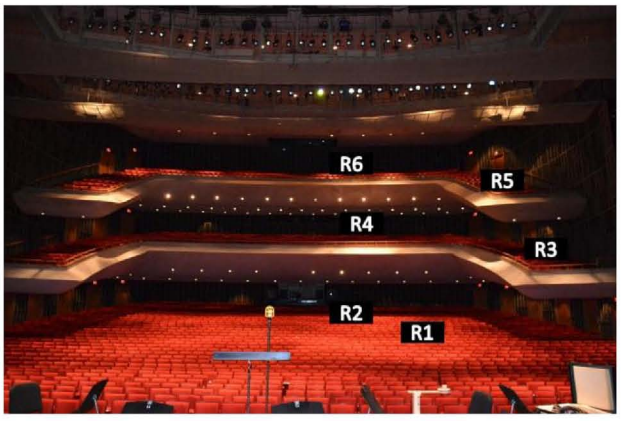


Figure 2: *Receiver positions in Eisenhower Auditorium.*

At each receiver position, the IR was measured using the measurement software EASERA. The stimulus was a swept sine signal with a pink-weighted spectrum. Each measurement was done using 10 sweep averages, and an additional pre-sweep. Measurements were taken using both microphone arrays: the Eigenmike and the omni and figure-8 pair. In terms of aligning the microphones with the sound source, the front of the Eigenmike array was pointed at the loudspeaker, while for the microphone pair, it was aligned with the null plane of the Sennheiser microphone oriented vertically toward the loudspeaker. The aligning of microphones toward the source was done by eye, which became increasingly more difficult for the receiver locations were towards the back of the hall.

To check for measurement repeatability, two additional sets of measurements were taken. In between sets, the custom microphone stand was removed and the various adjustment points were loosened and randomly repositioned. The stand was then replaced in the same spot to re-measure the IRs for each microphone. A total of three IR measurements were taken for each microphone array, at each of the six receiver locations for a total of 36 measurements.

3. DATA PROCESSING

3.1. Frequency Response Correction Filter

The raw frequency response of the Sennheiser MKH 30 microphone and Eigenmike array varied as much as ± 6 dB at certain frequencies, and this was deemed unacceptable for the measurements of LF and GLL. Filters were generated to compensate for the frequency response of both the Eigenmike array and the Sennheiser MKH 30 based on free field measurements.

A target for the filter magnitude as a function of frequency was generated by taking the difference between the on-axis magnitude of the Sennheiser MKH 30 frequency response and the on-axis magnitude of a B&K 4191 free field measurement microphone in dB. The difference was taken with one-third octave band logarithmic-energy-smoothing applied to both the figure-8 response and the measurement microphone response. The target was then fit to a minimum-phase FIR filter. A minimum-phase filter was used to keep the filter's IR compact in time and to avoid pre-ringing of the IR, which would occur in linear-phase or zero-phase filtering techniques. The same procedure was applied to the dipole pattern generated from the Eigenmike, although since the magnitude of each lobe on the figure-8 differs at low and high frequencies, an energy average of the magnitude of both lobes were used to create the target.

A similar method was used to create a filter for the omni-directional response of the Eigenmike. Instead of free-field measurements, the filter target was created by using the energy averaged (over receiver position) difference in dB between the B&K 4192 and the Eigenmike frequency responses in the hall.

3.2. Beamforming

To calculate LF and GLL from the Eigenmike measurements, the omni-directional response and dipole response must be extracted. These responses were generated using EigenStudio, a computer application by mh Acoustics for the Eigenmike, which performs the beamforming operation on the 32 channels of data. The 32-channel IRs recorded in EASERA were loaded into EigenStudio, which outputs both the omni-directional and dipole IRs, with the null plane oriented vertically toward the source. EigenStudio uses a two stage beamforming process [13]. In the first stage, the 32 channels are transformed into orthonormal beam patterns referred to as eigenbeams via a transformation to spherical harmonics. The second stage is a modal beamformer where each beam is weighted by a factor and the beams are summed to achieve the desired directivity.

3.3. Spatial Parameter Calculation

EASERA was used to calculate the spatial parameters LF and GLL. Using equation (1), LF was calculated for each octave band for the Eigenmike using the beamformed omni-directional IR and beamformed dipole IR for each of the three repetitions at each receiver location. In addition, LF was calculated for the omni and figure-8 pair using the IRs from the B&K 4192 and the Sennheiser MKH 30. The LF in the octave bands from 125 Hz to 1 kHz were arithmetically averaged together to obtain a single number for LF according to ISO 3382 [12]. Differences for LF are given in JNDs, where one JND is 0.05.

EASERA does not have a built in function to calculate GLL, so as an alternative, the Strength (G) function was adapted. In order to use this function, the figure-8 IR was used for the numerator of the strength calculation, with the first 80 milliseconds of the IR multiplied by zero. This modification was done using both the Eigenmike beamformed dipole IR and the Sennheiser MKH 30 IR. In both cases, the denominator for the GLL calculation was the anechoic response of the omni-directional source discussed in section 2.2. The GLL in the octave bands from 125 Hz to 1 kHz were energy-averaged together to obtain a single number for GLL according to ISO 3382 [12]. Differences for GLL are given in JNDs. The JND for GLL is not known, but for the purposes of this study the JND is assumed to be 1 dB, which is the JND for Strength (G).

4. RESULTS

4.1. Directivity

The frequency response function was measured every three degrees in the horizontal plane for each microphone. The magnitude was one-third octave band energy smoothed at each angle. For each microphone, polar plots of the magnitude were generated at each octave band. The Sennheiser MKH 30 directivity is very good broadband. As shown in Figure 3, the response is a consistent dipole over the entire frequency range of interest (125-4k Hz). The beamformed dipole response measured with the Eigenmike was very good for the 500 Hz to 4 kHz octave bands as shown in Figure 4. At 125 Hz, the response was not an ideal dipole. The magnitude of one of the lobes was smaller than the other lobe, and the nulls shifted in angle. This change occurs because at low frequencies EigenStudio sums in a portion of the zeroth order spherical harmonic, and the response approaches a cardioid pattern. While this technique is helpful in mitigating issues relating to white noise gain in certain applications, it is not ideal for measurements of LF and GLL.

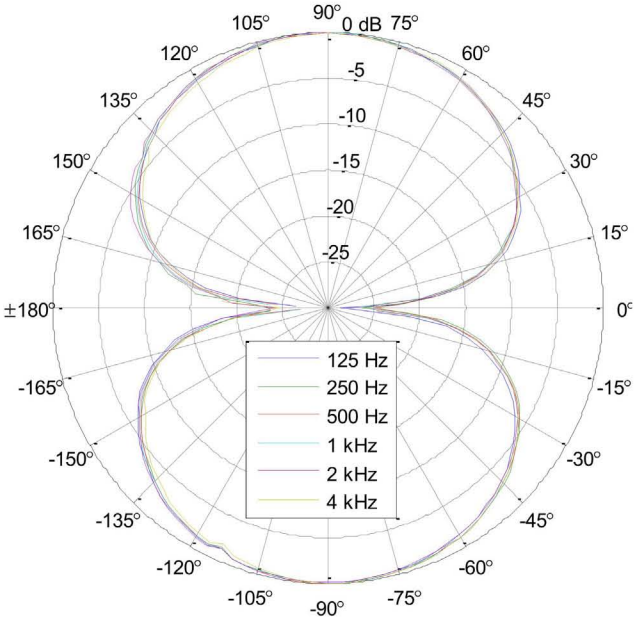


Figure 3: *Measured directivity pattern of the Sennheiser MKH 30.*

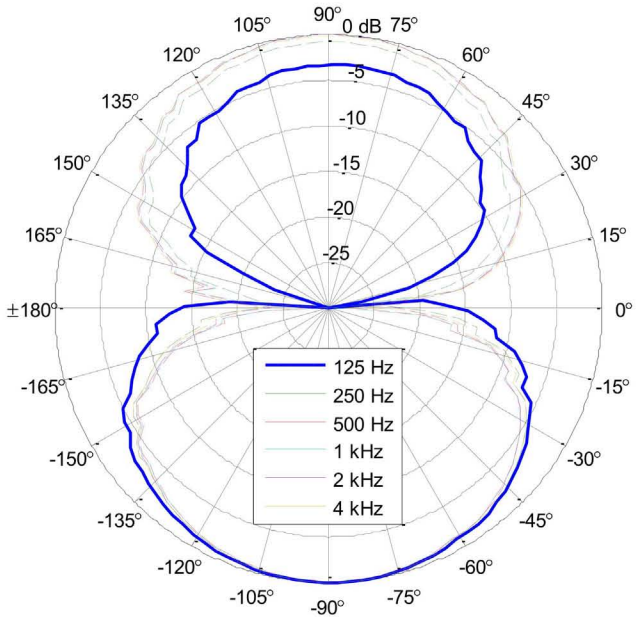


Figure 4: *Measured directivity of the Eigenmike dipole beamformed in Eigenstudio.*

4.2. Repeatability

The repeatability of the LF and GLL measurements for both the Eigenmike and omni-directional and figure-8 pair was assessed using the standard deviation of the three separate measurements taken at each of the six receiver locations. Since the number of measurements at each location was small, statistical tests were not performed. In terms of LF, repeatability was found to be good for R1 through R5 measured with the Eigenmike, which had standard deviations of 0.20 or less JNDs (see Table 1 for the average and Table 2 for the standard deviation). However, there was significantly more variation at R6, which had a standard deviation of 0.86 JNDs. This receiver position was in the back of the auditorium on the balcony, which was approximately 35 meters away from the stage, and was the most difficult to align the rotation of the microphone array. The repeatability was similar for the Sennheiser and B&K pair with lower variation at R2, R5, and R6, and slightly higher variation at R1, R3, and R4 (see Table 3 for the average and Table 4 for the standard deviation).

Table 1: *Average LF measured with Eigenmike.*

	125 Hz	250 Hz	500 Hz	1 kHz	Avg. 125-1k
R1	0.26	0.21	0.10	0.10	0.17
R2	0.12	0.15	0.08	0.09	0.11
R3	0.34	0.38	0.16	0.11	0.25
R4	0.19	0.16	0.11	0.15	0.15
R5	0.45	0.29	0.18	0.17	0.27
R6	0.16	0.16	0.28	0.28	0.22

Table 2: *Standard deviation of Eigenmike LF measurements in # of JNDs, where one JND is 0.05.*

	125 Hz	250 Hz	500 Hz	1 kHz	Avg. 125-1k
R1	0.04	0.07	0.03	0.05	0.05
R2	0.10	0.52	0.06	0.07	0.19
R3	0.15	0.07	0.10	0.01	0.08
R4	0.06	0.02	0.04	0.02	0.03
R5	0.18	0.23	0.20	0.16	0.20
R6	0.34	1.13	1.07	0.91	0.86

Table 3: *Average LF measured with Sennheiser and B&K pair.*

	125 Hz	250 Hz	500 Hz	1 kHz	Avg. 125-1k
R1	0.13	0.15	0.09	0.08	0.11
R2	0.05	0.10	0.07	0.09	0.08
R3	0.13	0.26	0.15	0.08	0.16
R4	0.12	0.13	0.10	0.14	0.12
R5	0.26	0.20	0.14	0.14	0.18
R6	0.12	0.24	0.19	0.19	0.18

Table 4: *Standard Deviation of Sennheiser and B&K pair LF measurements in # of JNDs, where one JND is 0.05.*

	125 Hz	250 Hz	500 Hz	1 kHz	Avg. 125-1k
R1	0.09	0.04	0.04	0.09	0.07
R2	0.13	0.07	0.05	0.07	0.08
R3	0.13	0.08	0.17	0.05	0.11
R4	0.08	0.07	0.02	0.07	0.06
R5	0.13	0.21	0.09	0.02	0.11
R6	0.20	1.52	0.15	0.11	0.50

The repeatability was found to be better for GLL than for LF in all receiver locations for the Eigenmike (see Table 5 for average GLL and Table 6 for standard deviation). This finding was true even in cases where the LF repeatability was relatively poor. These results indicate that GLL measurements are less sensitive to small spatial misalignments than LF measurements. The repeatability of the GLL measurements made using the Sennheiser and B&K pair were also similar to repeatability of the Eigenmike measurements (see Table 7 for average GLL and Table 8 for standard deviation).

Table 5: *Average GLL measured with Eigenmike [dB].*

	125 Hz	250 Hz	500 Hz	1 kHz	Avg. 125-1k
R1	-2.07	-0.93	-2.70	-4.03	-2.29
R2	-1.60	-3.03	-5.63	-5.27	-3.57
R3	-3.17	-2.67	-5.63	-4.40	-3.82
R4	-4.00	-3.93	-5.00	-4.53	-4.35
R5	-0.50	-1.53	-3.53	-3.67	-2.10
R6	-2.50	-1.93	-3.43	-3.87	-2.87

Table 6: *Standard deviation of Eigenmike GLL measurements in # of JNDs, where one JND is 1 dB.*

	125 Hz	250 Hz	500 Hz	1 kHz	Avg. 125-1k
R1	0.12	0.09	0.00	0.09	0.08
R2	0.22	0.12	0.12	0.12	0.15
R3	0.19	0.05	0.05	0.00	0.07
R4	0.08	0.05	0.00	0.05	0.04
R5	0.00	0.12	0.12	0.09	0.09
R6	0.16	0.05	0.09	0.05	0.09

Table 7: *Average GLL measured with Sennheiser and B&K pair [dB].*

	125 Hz	250 Hz	500 Hz	1 kHz	Avg. 125-1k
R1	-2.00	-0.90	-2.70	-4.20	-2.29
R2	-3.23	-3.23	-5.40	-5.33	-4.17
R3	-3.50	-2.20	-5.17	-4.33	-3.66
R4	-5.50	-3.63	-4.70	-4.53	-4.54
R5	-0.67	-1.47	-3.27	-3.50	-2.06
R6	-3.27	-2.00	-3.23	-3.93	-3.05

Table 8: *Standard Deviation of Sennheiser and B&K pair GLL measurements in # of JNDs, where one JND is 1 dB.*

	125 Hz	250 Hz	500 Hz	1 kHz	Avg. 125-1k
R1	0.00	0.00	0.00	0.00	0.00
R2	0.05	0.12	0.08	0.05	0.08
R3	0.08	0.08	0.05	0.05	0.06
R4	0.08	0.05	0.00	0.05	0.04
R5	0.05	0.05	0.09	0.00	0.05
R6	0.12	0.00	0.05	0.05	0.05

In IR measurements where repeatability was found to be poor for LF, the GLL measurements were largely unaffected. LF is calculated by integrating the lateral energy from 5 ms to 80 ms, which can be seen in equation (1). In this region, the microphone picks up energy from the direct arrival and early reflections. Ideally, the figure-8 microphone should reject both the direct sound and early reflections which are coming from directly in front of the microphone. Small angular misalignments will allow some of the direct sound in the measurement, and since the direct sound is high in level compared to later reflections, these misalignments could

have a significant impact on the measurement. This effect can be seen in the figure-8 IRs measured at R6. Figure 5 shows the three repetitions of the 250 Hz octave band measured with the Sennheiser microphone, which was the worst case for measurement repeatability. One of the measurements has more energy in the first 20 ms of the IR than the other two measurements, after which the IRs seem to agree more closely. Conversely, the GLL calculation involves integration of the lateral energy from 80 ms to infinity as seen in equation (2), and is less susceptible to small misalignments because there are no longer strong components which are directly on-axis in the late sound field.

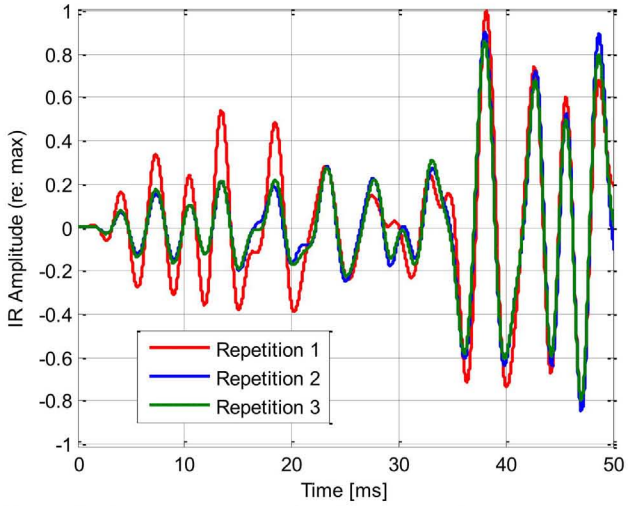


Figure 5: *Three repetitions of the Sennheiser IR measurement at R6 filtered at 250 Hz octave band.*

4.3. Two microphone comparison

The LF and GLL measurements made with the Eigenmike and Sennheiser MKH 30 were compared by subtracting one value from the other and then converted into the number of JNDS. The largest differences in LF between the two microphone configurations were found in the 125 Hz octave band, and to a lesser extent the 250 Hz octave band (see Table 9), where the LF measured with the Eigenmike is substantially higher than the LF measured with the Sennheiser MKH 30 in all receiver locations. This discrepancy is likely due to the fact that the null in the Eigenmike's dipole response beamformed in Eigenstudio has shifted in angle from the ideal dipole pattern. For these measurements, the null is pointed at the sound source on the stage. Since the null is shifted, the microphone is picking up portions of the direct sound and early reflections which are rejected in the Sennheiser measurement. An example of this effect can be seen in Figure 6, which shows the first 80 ms of the 125 Hz IRs for both the Eigenmike and Sennheiser microphones. Agreement between the Eigenmike and the Sennheiser is much better from 500 Hz to 4 kHz where the Eigenmike's directivity pattern is closer to an ideal dipole, with the exception of R6 where agreement was poor from 500 Hz to 4 kHz, which is also where there was relatively poor repeatability.

Table 9: *Difference in LF between Sennheiser and B&K pair and Eigenmike in # of JNDs, where one JND is 0.05.*

	125 Hz	**250 Hz**	**500 Hz**	**1 kHz**	**Avg. 125-1k**
R1	2.61	1.16	0.26	0.48	1.13
R2	1.49	1.09	0.19	0.07	0.71
R3	4.08	2.43	0.30	0.44	1.81
R4	1.52	0.57	0.24	0.02	0.59
R5	3.78	1.72	0.77	0.71	1.75
R6	0.85	1.57	1.93	1.82	0.76

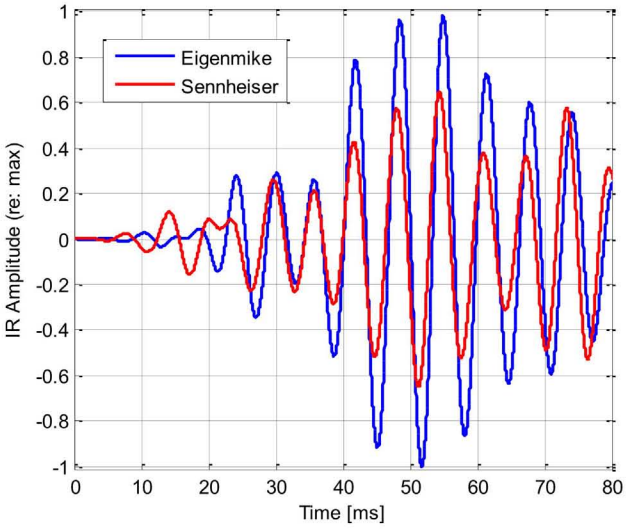


Figure 6: *Sennheiser and Eigenmike beamformed dipole IR measurements at R3 filtered at 125 Hz octave band.*

The measurement of GLL between the two microphones had better agreement than the measurement of LF. The largest variation is again seen in the 125 Hz octave band due to the non-ideal directivity of the beamformed dipole (see Table 10).

Table 10: *Difference in GLL between Sennheiser and B&K pair and Eigenmike in # of JNDs, where one JND is 1 dB.*

	125 Hz	**250 Hz**	**500 Hz**	**1 kHz**	**Avg. 125-1k**
R1	0.07	0.03	0.00	0.17	0.02
R2	1.63	0.20	0.23	0.07	0.48
R3	0.33	0.47	0.47	0.07	0.15
R4	1.50	0.30	0.30	0.00	0.29
R5	0.17	0.07	0.27	0.17	0.08
R6	0.77	0.07	0.20	0.07	0.19

5. CONCLUSIONS

The lateral energy component that is used to calculate the spatial parameters LF and GLL was measured using two different microphone configurations: a conventional figure-8 microphone (Sennheiser MKH 30), and a spherical microphone array (mh Acoustics em32 Eigenmike). Using the two methods, room impulse responses were measured in a 2500 seat auditorium and three repeatability measurements were taken in all six of the receiver locations.

The repeatability was evaluated for the spatial measures of LF and GLL. The LF measurement repeatability was found to be poor at R6, which was the location farthest from the stage. The low frequency standard deviation was more than one JND in some octave bands at R6. A likely cause for the repeatability error was misalignment in the rotation of the microphone, which would allow the figure-8 microphone to pick up a portion of the direct sound and early on-axis reflections, skewing the measurement of LF The measurements of GLL were much more consistent than the measurements of LF.

The averages of the three measurements at each receiver location were compared for each method. The difference of the LF average from the 125 Hz to 1 kHz octave bands for the two microphone configurations was between 0.59 and 1.81 JNDs at the six receiver locations. The largest differences were found in the 125 Hz and 250 Hz octave bands where the Eigenmike's dipole directivity is not ideal, while the differences were relatively small from 500 Hz to 4 kHz. The difference of the GLL average from 125 Hz to 1 kHz for the two methods was between 0.02 and 0.48 JNDs, with the largest variation in the 125 Hz octave band, which is most likely caused by the Eigenmike's low frequency non-ideal directivity pattern.

Future work will include manually calculating the spherical harmonic expansion of the Eigenmike's impulse responses to compare to the results from the included software, Eigenstudio. Since the LF and GLL agreement between the two methods was better at frequencies where the Eigenmike had an optimal directivity pattern, it is very likely that a manual calculation of the dipole component will yield better agreement with the Sennheiser microphone. Future work could also include simulations of changes in LF and GLL with variations in receiver directivity pattern and small spatial misalignments.

6. ACKNOWLEDGMENTS

The authors wish to express their thanks to Mr. Tom Hesketh, for allowing the authors access to the auditorium for taking the measurements. The authors also wish to acknowledge Matthew Neal, Matt Kamrath, Martin Lawless, and Acadia Kocher for their assistance with the measurements. The authors would also like to thank Bose Corporation in Framingham, MA for use of their anechoic chamber.

This work was sponsored by NSF award #1302741.

7. REFERENCES

[1] T. Lokki, J. Patynen, A. Kuusinen, and S. Tervo, "Disentangling preference ratings of concert hall acoustics using subjective sensory profiles," *J. Acoust. Soc. Am.*, vol. 132, no. 5, pp. 3148-3158, November 2012

[2] M. Barron, M., "Subjective effects of first reflections in concert halls- the need for lateral reflections," *J. Sound & Vib.*, vol. 15, no. 4, pp. 475-494, April 1971.

[3] M. Barron, and A. H. Marshall, "Spatial impression due to early lateral reflections in concert halls: The derivation of a physical measure," *J. Sound & Vib.*, vol. 77, no. 2, pp. 211-232, July 1981

[4] M. Morimoto, and Z. Maekawa, "Effects of low frequency components on auditory spaciousness," *Acustica*, vol. 66, no. 4, pp. 190-196, September 1988

[5] J. S. Bradley, and G. A. Soulodre, "The influence of late arriving energy on spatial impression," *J. Acoust. Soc. Am.*, vol. 97, no. 4, pp. 2263-2271, April 1995

[6] J. S. Bradley, and G. A. Soulodre, "Objective measures of listener envelopment," *J. Acoust. Soc. Am.,* vol. 98, no. 5, pp. 2590-2597, November 1995

[7] Lundeby, A., Vigran, T.E., Bietz, H., and Vorländer, M., "Uncertainties of measurements in room acoustics," *Acoustica*, vol. 81, no. 4, pp. 344-355., July/August 1995

[8] I. Witew, G. K. Behler, and M. Vorländer, "Spatial variation of lateral measures in different concert halls," *Proc. Int. Cong. on Acoust.*, Kyoto, Japan, pp. 2949-2952, 2004

[9] I. Bork, "Report on the 3rd round robin on room acoustical computer simulation – Part I: Measurements," *Acta Acustica with Acustica*, vol. 91, no. 4, pp. 740-752., July/August 2005

[10] M. C. Vigeant, C. B. Giacomoni, and A. C. Scherma, "Repeatability of spatial measures using figure-of-eight microphones," *Applied Acoustics*, vol. 74, , no. 9, pp. 1076-1084, September 2013

[11] mh acoustics, "em32 Eigenmike microphone array release notes (v17.0)", [online] 2013, http://www.mhacoustics.com/sites/default/files/ReleaseNotes.pdf (Accessed: 22 November 2013).

[12] *Acoustics — Measurement of room acoustic parameters — Part 1: Performance spaces*, ISO Standard 3382-1:2009(E)

[13] G. W. Elko, J. Meyer, "Microphone Arrays," in *Springer Handbook of Speech Processing*, Benesty, Sondhi, Huang, Springer Berlin Heidelberg, 2008, ch. 50, sec. 50.5, pp. 1034-1037

IMPLEMENTATION AND PERCEPTUAL EVALUATION OF A SIMULATION METHOD FOR COUPLED ROOMS IN HIGHER ORDER AMBISONICS

Giso Grimm,
Medical Physics
Universität Oldenburg and
Cluster of Excellence "Hearing4all",
Oldenburg, Germany
g.grimm@uni-oldenburg.de

Torben Wendt,
Acoustics and Medical Physics
Universität Oldenburg and
Cluster of Excellence "Hearing4all",
Oldenburg, Germany
torben.wendt@uni-oldenburg.de

Volker Hohmann,
Medical Physics
Universität Oldenburg and
Cluster of Excellence "Hearing4all",
Oldenburg, Germany
volker.hohmann@uni-oldenburg.de

Stephan D. Ewert,
Medical Physics
Universität Oldenburg and
Cluster of Excellence "Hearing4all",
Oldenburg, Germany
stephan.ewert@uni-oldenburg.de

ABSTRACT

A fast and perceptively plausible method for rendering acoustic scenarios with moving sources and moving listeners is presented. The method is principally suited for application in dynamic and interactive evaluation environments (e.g., for hearing aid development), psycho-physics with adaptively changing the spatial configuration, or simulation and computer games. The simulation distinguishes between the direct sound, sound reflected and diffracted by objects of limited size, diffuse sound surrounding the listener, e.g., diffuse background sounds and diffuse reverberation, and 'radiating holes' for simulation of coupled adjacent rooms. Instead of providing its own simulation of room reverberation, the proposed simulation method generates appropriate output signals for external room reverberation simulators (e.g., see contribution by Wendt et al.). The output of such room reverberation simulators is then taken either as diffuse surrounding sound if the listener position is within the simulated room, or as input into a 'radiating hole', if the listener is in an adjacent room.

Subjective evaluations are performed by comparing measured and synthesized transitions between coupled rooms.

1. INTRODUCTION

A large variety of tools for acoustic simulation of rooms and open spaces exists (e.g., [1], [2]). Most of these tools can simulate room acoustics at a very high accuracy. However, the required high complexity allows only to simulate static impulse responses or soundscapes. The time-varying simulation of acoustic spaces in real-time implies strong simplifications to typically used approaches of geometric acoustics (e.g., image sources [3]) and ray tracing ([4]) as well as combinations (e.g., [5]). Alternatively, pre-rendered acoustic scenes can be used (e. g. [6]). Dramatic simplifications are typical in real-time applications like computer games which seek immersive soundscapes at cost of realism. In this case often simple reverberation algorithms are used with settings predefined by the game designer. However, in opposite to static, physically accurate room simulation plausibility can, for specific applications, be more important than exact correspondence to the real world. Whereas a plausible approximation of room acoustics is often possible by the use of rectangular "shoebox" rooms for early reflection simulation in combination with a model for diffuse reverberation, the simulation of connected (coupled) rooms is more computationally demanding, nevertheless required for many auralization applications.

Acoustic simulations can be physically assessed by measures related to the simulated binaural room impulse response (e.g., decay time, early decay time, interaural cross correlation). Especially the simulation of existing environments can be compared with recorded impulse responses. Plausibility of acoustic simulations, however, is a more vague criterion. It may be measured by the perceptual component of auditory spatial awareness [7]: If a listener is able to identify a simulated environment purely by its acoustics, then a simulation may be seen as plausible.

In the present paper, the room coupling model of a toolbox for dynamic real-time simulation of acoustic spaces for hearing research and hearing aid evaluation (toolbox for acoustic scene creation and rendering, TASCAR, [8]) is described and perceptually evaluated. The toolbox aims to provide a physically accurate simulation of the direct sound and the first order reflections, together with a plausible reproduction of diffuse reverberation, diffuse environmental sounds and room coupling.

In Section 2, the simulation methods used by the toolbox are briefly described. Section 3 describes the evaluation methods used in this study and the measurement of real-world binaural impulse response that are used as a reference. The results (Section 4) are discussed and summarized.

2. SIMULATION METHODS

The proposed simulation method for dynamic transitions between coupled rooms is part of the TASCAR toolbox for acoustic scene creation and rendering [8]. The focus of this toolbox is the time-

varying simulation of acoustics, i.e., all sound sources, receivers and reflecting objects can move dynamically, and the simulated acoustics is returned as a time signal, depending on the input signals and the time-varying spatial configuration of the simulated environment. In this framework, point sources follow a distance model with a $1/r$ sound pressure law, r being the distance between sound source and receiver. Additionally, air absorption is approximated by a simple first order low-pass filter model:

$$y_k = a_1 y_{k-1} + (1 - a_1) x_k \quad (1)$$

$$a_1 = e^{-\frac{r f_s}{c \alpha}}, \quad (2)$$

where c is the speed of sound. The empiric constant $\alpha = 7782$ was manually adjusted to provide sensible values for distances below 50 meters. The resulting absorption as a function of distance is given in Figure 1. This approach is very similar to that of [9] who used a FIR filter to model the frequency response at certain distances. However, in this approach the distance parameter r can be varied dynamically.

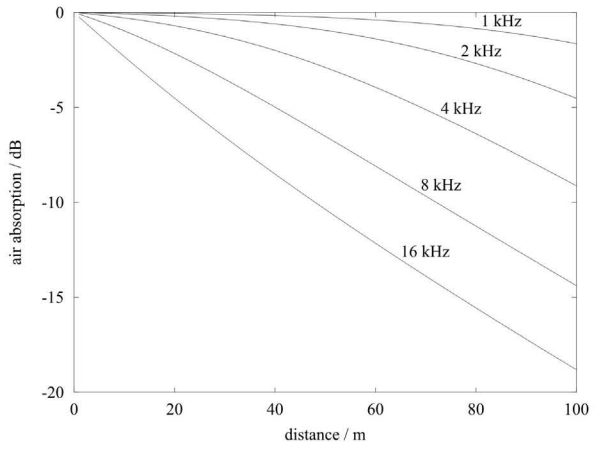


Figure 1: *Simulated air absorption as a function of distance. For distances below 50 m the simulated absorption in dB is roughly proportional to the distance. For larger distances, the low frequency absorption is too high.*

Early reflections are modeled using a image source model. In opposite to most commonly used models (e.g., [3]) which calculate impulse responses for a rectangular enclosure ("shoebox model"), reflections are simulated for each reflecting (rectangular, but arbitrarily oriented) surface. Diffraction is simulated by an empirically found attenuation g, which depends on the position of the image source $\mathbf{x}_{src}$, the receiver position $\mathbf{x}_{rec}$, and the point on the reflecting surface with the shortest connection between the image source and the receiver $\mathbf{x}_c$:

$$g = \left((\mathbf{x}_{src} - \mathbf{x}_c) \cdot (\mathbf{x}_c - \mathbf{x}_{rec})\right)^{2.7} \quad (3)$$

In this study, only reflections of first order are simulated. Primary sources and reflected sources are simulated as omnidirectional sources.

Diffuse sources, e.g., background signals, or diffuse reverberation, are added in first order ambisonics (FOA) format. No distance law is applied to diffuse sound sources; instead, they have a rectangular spatial range box, i.e., they are only rendered if the receiver is within their range box, with a von-Hann ramp at the boundaries of the range box. Position and orientation of the range box can vary with time. The diffuse source signal is rotated by the difference between receiver orientation and box orientation.

In the acoustic simulation, receivers can be considered as virtual microphones, i.e., receivers return the output signals of the acoustic simulation. Two receiver types are used: A *spatial receiver* encodes the direction of the direct and reflected sound sources in 3rd order ambisonics, and adds the diffuse sound sources to the first order components of the receiver output[1]. *Omnidirectional receivers* with a single output are used to return the signal of the direct and reflected sources to the external diffuse reverberation generators. Both receiver types apply the distance and air-absorption mode to the source signals. The omnidirectional receivers can have a finite range box. If a source is within that range box of a receiver the distance law is only applied to the delay and not to the gain and air absorption model. This way it can be achieved to have all sources within a simulated room contributing with the same gain to the diffuse reverberation. Both receiver types can be restricted to render only sources within a range box, and they can also be restricted to render only direct point sources, mirrored point sources or diffuse sources.

In the proposed simulation framework, each room is simulated separately. Coupling is provided by 'sound portals' which are situated in the (door) openings connecting adjacent rooms. A sound portal has a virtual sound source attached radiating from the nearest point of the surface of the portal surface to the receiver. In the distance law, however, the distance between the sound portal and the source room center is added, to create a plausible fall-off rate. The sound radiated from the sound portal is gained from an omnidirectional receiver in the source room behind the portal surface.

Diffuse reverberation is rendered by an external tool. Each simulated room provides an omnidirectional (or optionally a directional) receiver. The output signal of that receiver is processed by a diffuse reverberation model [10] configured to match the specific room qualities. The first order ambisonics reverberation signal is then added as a diffuse input source.

3. EVALUATION METHODS

3.1. Simulated environment

The simulated environment was an office room ($4.43 \cdot 4.5 \cdot 3$ m^3) next to a long corridor ($30 \cdot 1.94 \cdot 2.5$ m^3). The sound source was always in the office room at a fixed position. Five static listening positions (see Figure 2) have been rendered: The first was close to the source in the office room, facing towards the sound source. The second position was closer to the door, but facing the same direction. The third listening position was exactly in the door between the office room and the corridor. The fourth and fifth position were in the corridor, with the source hidden by the wall. In the dynamic situations the position was interpolated linearly, within 20 seconds, resulting in a velocity between 0.14 and 0.35 m/s.

In the room simulation model, reflectors were placed at all walls, with the exception of the door. In the office, the direct sound

[1]The authors are aware of artifacts caused by playback of first order sources via higher order systems, caused by near field compensation order weights and coloration due to correlated sources played over many loudspeakers. However, both limitations do not play a major role in this setup, since first order sources are only used for diffuse and thus uncorrelated sounds, and near field compensation is not applied here.

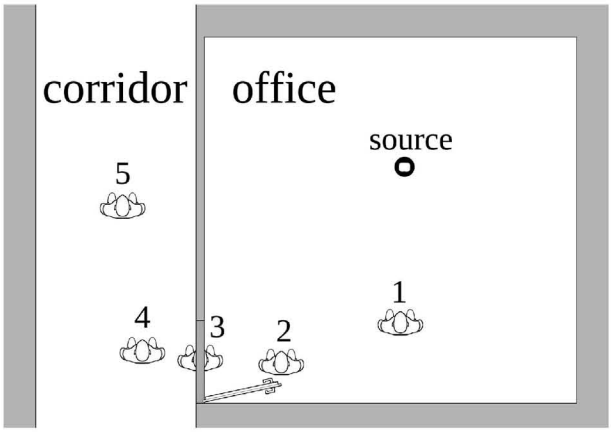


Figure 2: *Simulated environment, with listening positions marked by an artificial head.*

Position	1	2	3	4	5
Distance [m]	1.87	2.78	3.34	4.04	5.40
Azimuth [deg.]	-1.5	-32.4	-46.5	-97.3	-152.9

Table 1: Distance and direction between receiver and sound source. For positions 4 and 5, the distance is the sum of the distance receiver-sound portal and sound portal-source. The azimuth of positions 4 and 5 represents the direction to the sound portal.

and the early reflections were recorded by an omnidirectional receiver in the size of the office room. The output of this receiver was used as an input of the diffuse reverberation simulator. The output of the diffuse reverberation simulator was added to the listener when in the office room, and fed into the 'door' input, which was propagating it further into the corridor. The 'door' output was used as a directed sound by the listener receiver, and used as input of the diffuse reverberation simulator of the corridor. The von-Hann ramps of the receivers were 0.2 m long.

In Table 1 the distance and direction between receiver and source is given. For the positions 4 and 5, the sum of the distance receiver-sound portal and sound portal-source is given, as well as the direction between the receiver and the door.

3.2. Apparatus and test stimuli

All signals used in the subjective evaluation were pre-rendered. Stimuli were played back via a Sennheiser HD580 headphone. The levels could be adapted individually to match a comfortable level during the beginning of the measurement, and were kept fix during the measurement.

The receiver used for listening was a full periphonic 3rd-order ambisonics encoder. The receiver output was routed to an ambisonics decoder [11] to get the loudspeaker feeds. Binaural signals were generated by convolution with binaural head-related impulse responses (HRIR) of an artificial head for the 20 virtual speaker positions on the vertices of a regular dodecahedron. In the situation with static receiver (listener) positions binaural room impulse responses (BRIR) were generated by using a Dirac-delta pulse as the input signal of the primary source. Two anechoic test signals with a duration of 20 seconds, a speech signal (labeled 'speech') and a music recording [12] (labeled 'guit.'), were convolved with the BRIRs. As a reference condition, BRIRs have been recorded in the equivalent real room. Room impulse responses were measured using an omnidirectional loudspeaker. For recording, an artificial head MK2 by Cortex with a respective measurement amplifier Manikin MK1 was used. The excitation signal was a logarithmic sweep [13], with starting and ending frequencies of 50 Hz and 18 kHz, respectively. See [10] for details. The same measurement procedure and equipment was used to obtain the HRIR database for binaural auralization.

The test signals with a moving listener are directly rendered using the same anechoic source input signals. The time-varying listening position was linearly interpolated between the five discrete positions. Additionally, for the dynamic situation, a visual simulation of the scenario was rendered as a video using a 3d-graphics tool [14], with the same spatial properties of the simulated environment as in the acoustic simulation.

position	1	2	3	4	5
simulated:					
L_l [dB]	-2.7	-10.2	-11.7	-11.3	(-15.3)
L_r [dB]	-1.9	-5.7	-1.4	9.6	(4.5)
$T_{60,1}$ [s]	0.16	0.19	0.13	0.02	0.09
$T_{60,2}$ [s]	0.42	0.43	0.48	0.59	0.60
ILD [dB]	-0.1	0.1	-1.1	0.4	-4.1
IACC	0.52	0.21	0.17	0.12	0.13
recorded:					
L_l [dB]	-4.2	-11.9	-11.3	-14.2	(-6.1)
L_r [dB]	-5.8	-5.0	-5.4	-3.7	(-12.8)
$T_{60,1}$ [s]	0.25	0.27	0.33	0.42	0.40
$T_{60,2}$ [s]	0.37	0.42	0.47	0.67	0.65
ILD [dB]	1.0	4.2	7.3	5.9	-1.9
IACC	0.32	0.08	0.13	0.08	0.09

Table 2: Physical characterization of the BRIR at the tested listening positions. L_l and L_r denote the 'liveness' [15] at the left and right ear, respectively. None of the positions 1 to 3 are within the critical distance. $T_{60,1}$ and $T_{60,2}$ are the early and late reverberation times [16]. The early reverberation time is lower for the simulated BRIRs than for the recorded BRIRs. Interaural level difference (ILD) suggest a small lateralization for the simulated method opposed to a large lateralization for the recorded BRIR. The interaural cross correlation (IACC) is slightly higher in the simulation. In the listening position 5 no direct sound was audible, thus the 'liveness' measures L_l and L_r have to be interpreted with care and are given in brackets.

3.3. Physical characterization of the BRIR

To characterize the simulated and recorded BRIR, the 'liveness' L [15], i.e., the ratio between the direct sound to the reverberant sound, has been derived from the impulse responses. This is related to the critical distance, at which L equals 0, i.e., values of $L \geq 0$ represent a receiver within the critical distance. Additionally, the early and late reverberation time, $T_{60,1}$ and $T_{60,2}$, measured after [16], are given for comparison. Interaural level difference (ILD) and the interaural cross correlation (IACC, [17]) as used by [10] are also provided. The data are given in Table 2.

3.4. Spatial awareness

To evaluate the spatial quality of the simulated room coupling, the auditory spatial awareness [7] was measured using an identification task: In a graphical interface, the listener was able to switch between the stimuli for the five listening positions, without knowing the order. The order of the stimuli was randomized. They were time-aligned and repeated, i.e., by switching between the stimuli only the spatial configuration of the sound changed. The task was to assign the number of the listening position to each sound. The users were able to revise their decision at any time; when a listening position was assigned to all sounds the subjects were able to finally confirm their decision and end the measurement. A confusion matrix was calculated from the results. If the acoustic stimulus delivered a spatial awareness, then only the diagonal elements would differ from zero. As a measure of spatial quality, the accuracy, i.e., the sum of correctly identified stimuli divided by the sum of presented stimuli, was calculated. The individual accuracy for each stimulus and presentation method was calculated as well as the pooled accuracy for all subjects and test stimuli in each presentation method. Additionally to the accuracy, specific confusions bear the potential of revealing artifacts of the simulation method.

The spatial awareness measurement on a ranked path, as applied here, can be regarded as an indirect measure of the combination of source localization and distance perception.

Both, recorded BRIR and simulated BRIR, were tested.

3.5. Subjective rating

Absolute rating of naturalness was asked for during the playback of the simulated video in combination with the simulated sounds, for the speech signal and the music signal. The subjects were ask to rate the overall naturalness, the naturalness in the office room, in the corridor, and during the transition. The listeners knew the simulated position by watching the video representation.

3.6. Test subjects

Thirteen normal hearing listeners (average age 31 years, standard deviation 7 years) participated in the experiment.

4. RESULTS

Results of the effect of the simulation method on spatial awareness are shown in Fig. 3 and 4. Figure 3 shows the confusion matrix between presented and perceived listening position for simulated static room impulse responses (upper panels) and measured binaural room impulse responses (lower panels), for the two test stimuli.

All positions were detected correctly by more than 50% of the test subjects, for any reproduction method and test signal. However, with the simulation method there is a confusion noticeable between position 3 (door) and 4 (corridor, close to door). This is also represented by the decreased accuracy (median of the individual accuracy of 0.6 for simulated acoustics versus 1 for recorded BRIR, or pooled accuracy over all test signals and subjects of 0.68 for simulated acoustics versus 0.88 for the recorded BRIR). Figure 4 shows median values, quartile-ranges and extrema of the accuracy for the different test conditions. A two-way analyis of variance (ANOVA) revealed that the effect of the presentation method (simulation or BRIR) on accuracy is statistically significant (p=0.05); the test stimulus has no significant effect.

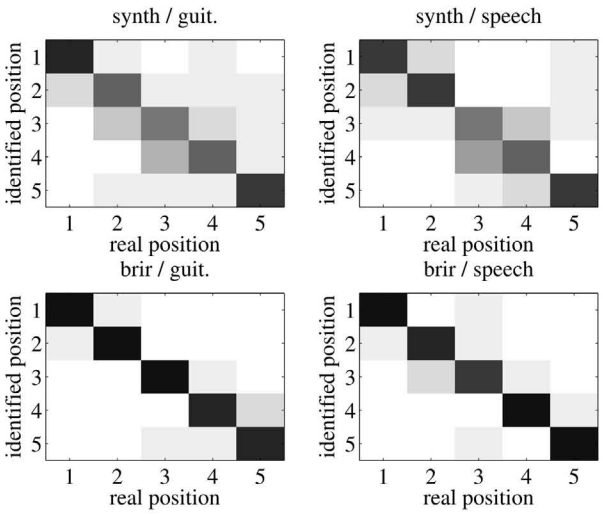


Figure 3: *Confusion matrices for the four tested conditions; results for simulated listening positions are shown in the upper panels, and for recorded BRIRs in the lower panels. For the simulated listening positions, a confusion between listening position 3 and 4 is noticeable.*

Absolute rating of naturalness (Figure 5) in the dynamic condition shows a moderate to high naturalness. The naturalness is perceived best in the office, i.e., when the direct sound of the source is audible. The 'corridor' listening position is rated slightly worse, which was also supported by some test subject comments mentioning that the sound level in the corridor was perceived to high compared to the sound level in the office. A two-way ANOVA with factors listening position and test stimulus [4x2] revealed a significant main effect of listening position (p=0.05). No significant effect of test stimulus was observed. Post-hoc comparison using Fisher's LSD criterion reveals that the rating in the office is significantly better than in all other tested conditions (p=0.05).

5. DISCUSSION

The evaluation of the simulation method for coupled rooms shows deficits when being compared to measured binaural room impulse responses. While extreme positions were correctly identified in most cases, it was more difficult to distinguish between the listening positions close to the door. This might be caused by the fact that the door simulation source radiates the sound of the office room with an insufficient falloff rate, and that the direct sound is not disappearing appropriately when moving from listening position 3 to position 4. It is primarily this confusion which decreases the accuracy to values around 0.6. The confusion between the first and the second listening positions might be a hint that the first early reflections are to strong, and that higher order reflections are needed for a better distinction between the listening positions within the room. This is also supported by the early reverberation time, which is smaller for the simulation than for recorded BRIR, and by the interaural cross correlation, which is higher for the simulation. Further more, the physically measured ILD does not reflect the lateralization of the source when moving from position 1 to position 3. Also a coupling from the corridor back to the

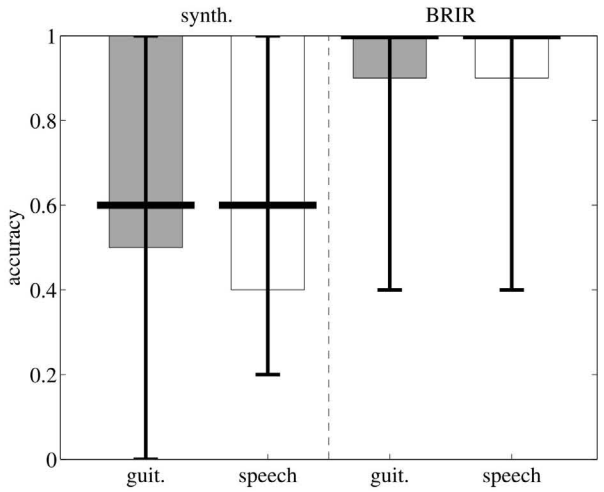


Figure 4: *The accuracy of the simulated (left half) and recorded (right half) listening positions. Median values (thick black lines), inter-quartile ranges (gray bars) and extrema across all test subjects are shown. An accuracy of 0.6 is reached if a single pair of listening positions is exchanged, i.e., most subjects were able to identify three listening positions correctly with the simulated listening positions, and identified all listening positions correctly with the recorded BRIR.*

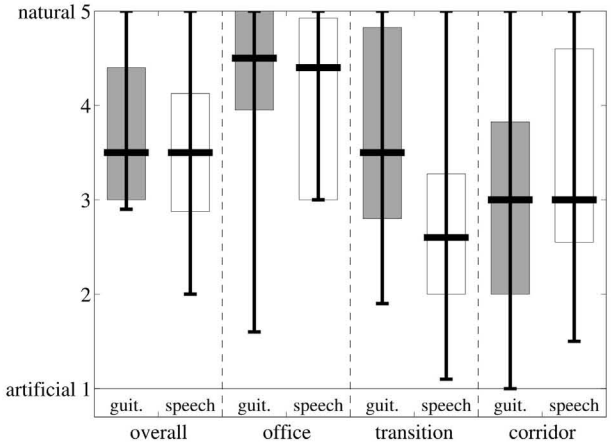


Figure 5: *Rating of the naturalness of the dynamic simulation. A video with a visual simulation of the environment from the listener's perspective was presented together with the simulated audio for either a music test signal ('guit.') or a speech test signal ('speech'). The subjects were ask to rate the overall naturalness and the naturalness of the simulation in the office, during the transition through the door, and in the corridor. Naturalness was rated on a five-point scale.*

office room, which was not applied here, may improve the spatial awareness.

In general the method of measuring the spatial quality in terms of spatial awareness is suitable to show differences between simulation methods, even if the stimuli differ in their spectra. A direct comparison is likely to be dominated by the spectral differences of the stimuli; this effect can be excluded using the spatial awareness test. On the other hand, spatial awareness is only one of many facets of spatial reproduction methods; thus this test method can not supersede other tests. Specifically, the presented method can only be applied to static positions, whereas the other test method of this study, absolute rating of naturalness, can also be applied to dynamic stimuli.

Some test subjects commented on the dynamic simulation after the experiment: Three subjects mentioned that the overall level of the test signal in the corridor was to high in relation to the direct sound in the office. High values of the 'liveness' measure for the simulation in the corridor may indicate that the virtual source in the door was too dominant. This level difference may have been the reason for the slightly lower naturalness rating in the corridor. Also some test subjects remarked that the transition between the rooms was difficult to rate for the speech signal, because there was a gap between sentences exactly in the door. This is likely to be responsible for the lower rating of the speech signal in the transition.

The current evaluation showed satisfying plausible reproduction for the given path. Future research will investigate whether the proposed subjective evaluation method and the BRIR synthesis method are more generally applicable to arbitrary paths in different rooms.

6. CONCLUSIONS

A computationally highly efficient real-time simulation method for transitions between coupled rooms was presented. The performance of the simulation method has been evaluated by measures of spatial awareness and by absolute subjective ratings. Although in absolute ratings the overall naturalness is rated high, the spatial awareness test reveals specific problems in the suggested simulation method. The presented data proof the function and applicability of the method and provide valuable input for further improvement of the simulation method. An improved version should better reproduce the ILD and the early reverberation time. The ILD could be improved by attenuating the diffuse sources in the direction of nearby walls. The early reverberation time can be improved by adding higher order reflections in the mirror source model. The unnaturally high amount of direct sound at the listening positions in the corridor can be improved by excluding the direct sound in the sound radiated by the sound portal in the door opening. A more systematic evaluation would require to investigate more paths, and different rooms.

7. ACKNOWLEDGMENTS

This study was funded by the DFG FOR 1732 "Individualisierte Hörakustik".

8. REFERENCES

[1] "Ease," http://ease.afmg.eu/.

[2] "odeon," http://www.odeon.dk/.

[3] Jont B Allen and David A Berkley, "Image method for efficiently simulating small-room acoustics," *The Journal of the Acoustical Society of America*, vol. 65, pp. 943, 1979.

[4] A. Krokstad, S Strøm, and S. Sørsdal, "Calculating the acoustical room impulse response by the use of a ray tracing technique," *J. Sound Vib.*, vol. 8, no. 1, pp. 118–125, 1968.

[5] Michael Vorländer, "Simulation of the transient and steadystate sound propagation in rooms using a new combined ray-tracing/image-source algorithm," *J. Acoust. Soc. Am.*, vol. 86, pp. 172–178, 1989.

[6] B.-I. Dalenbäck and M. Strömberg, "Real Time Walkthrough Auralization – The First Year," Tech. Rep., CATT (Dalenbäck), Valeo Graphics (Strömberg), 2010.

[7] Barry Blesser and Linda-Ruth Salter, *Spaces Speak, Are You Listening?*, The MIT Press, 2007.

[8] Giso Grimm, Graham Coleman, and Volker Hohmann, "Realistic spatially complex acoustic scenes for space-aware hearing aids and computational acoustic scene analysis," in *16. Jahrestagung der Deutschen Gesellschaft für Audiologie*, Rostock, Germany, 2013.

[9] Jyri Huopaniemi, Lauri Savioja, and Matti Karjalainen, "Modeling of reflections and air absorption in acoustical spaces a digital filter design approach," in *Applications of Signal Processing to Audio and Acoustics, 1997. 1997 IEEE ASSP Workshop on*. IEEE, 1997, pp. 4–pp.

[10] Torben Wendt, Steven van der Par, and Stephan D. Ewert, "Perceptual and room acoustical evaluation of a computational efficient binaural room impulse response simulation method," in *Proceedings of the EAA Joint Symposium on Auralization and Ambisonics*, 2014, submitted.

[11] Fons Adriaensen, "Ambdec 0.5.1 - first, seond and third order Ambisonics decoder," `http://kokkinizita.linuxaudio.org/linuxaudio/`, 2011.

[12] Michio Woirgardt, Philipp Stade, Jeffrey Amankwor, Benjamin Bernschütz, and Johannes Arend, "Cologne university of applied sciences – anechoic recordings," `http://www.audiogroup.web.fh-koeln.de/`, 2012.

[13] Angelo Farina, "Simultaneous measurement of impulse response and distortion with a swept-sine technique," in *Audio Engineering Society Convention 108*, 2 2000.

[14] "blender," `http://www.blender.org/`, 2013, Stichting Blender Foundation, Amsterdam, the Netherlands.

[15] JP Maxfield and WJ Albersheim, "An acoustic constant of enclosed spaces correlatable with their apparent liveness," *The Journal of the Acoustical Society of America*, vol. 19, pp. 71–79, 1947.

[16] Manfred R Schroeder, "New method of measuring reverberation time," *The Journal of the Acoustical Society of America*, vol. 37, pp. 409–412, 1965.

[17] T. Hidaka, L. L. Beranek, and T. Okano, "Interaural cross-correlation, lateral fraction, and low- and high-frequency sound levels as measures of acoustical quality in concert halls," *The Journal of the Acoustical Society of America*, vol. 98, pp. 988–1007, 1995.

EVALUATION OF AMBISONICS DECODING METHODS WITH EXPERIMENTAL MEASUREMENTS

Diego M. Murillo[1], Filippo M. Fazi[2], Mincheol Shin[3]

[1]Institute of Sound and Vibration,
University of Southampton
Southampton, United Kingdom
dmmg1c12@soton.ac.uk

[2] Institute of Sound and Vibration,
University of Southampton
Southampton, United Kingdom
ff1@isvr.soton.ac.uk

[3] Institute of Sound and Vibration,
University of Southampton
Southampton, United Kingdom
M.Shin@soton.ac.uk

ABSTRACT

Ambisonics is a sound reproduction technique based on the decomposition of the sound field using spherical harmonics. The truncation in the number of coefficients used to recreate the sound field leads to reproduction artifacts which depend on the frequency and the listener spatial location. In this work, the performance of three different decoding methods (Basic, Max-rE and In-Phase) has been studied and evaluated in the light of the results of experimental measurements. The latter were performed using a spherical array composed of 40 uniformly distributed loudspeakers and a translating 29-channel linear microphone array. An error analysis is presented based on the difference between the desired and synthesized sound pressure and acoustic intensity field. The results indicate that, as expected, the size of the region of accurate sound field reconstruction reduces as frequency increases, but with different trends depending on the type of decoder implemented.

1. INTRODUCTION

3-D audio reproduction allows the generation of virtual spaces where the user perceives the sound according to the acoustic characteristics of the environment. This immersive experience has wide applications in areas such an entertainment, education, and research, among others. One methodology commonly used to reconstruct 3-D sound is the use of multichannel systems that reproduce the desired sound field over a specific area. Some advantages of implementing these techniques are a better immersive experience due to the use of multiple loudspeakers and the fact that the listening cues as Interaural Time, Level and Phase Differences are created in a natural way by the listener [1].

Ambisonics is a multichannel technique which has been extensively applied from the seventies [2-4]. It is based on the decomposition of the sound field using spherical harmonics which are part of the solution of the wave equation when it is expressed in spherical coordinates [5]. In theory, an exact reconstruction of the sound field is given when an infinite number of coefficients are computed. However, if this number is finite, the truncation will decrease the accuracy of the reproduction, depending on the frequency and the spatial location. The selection of the order of spherical harmonics is determined by the number of loudspeakers available for the reproduction of the sound field. This relation is commonly expressed by the following rule of thumb [6]:

$$L \geq (kr+1)^2 \tag{1}$$

where L is the number of loudspeakers, k is the wavenumber and r is the radius of the area where the reconstruction is accurate (radius of validity or reference radius). Equation 1 implies that a high number of loudspeakers when the reproduction of high frequency sound is attempted over a wide area[1], generating a trade-off between these two variables.

Due to the artifacts created by the truncation in the number of spherical harmonics, different methods have been proposed to increase the physical or perceptual accuracy of the sound field reconstruction. For example, Max-rE decoder aims to maximize the energy vector optimizating the high frequency sound reproduction. The energy vector is defined as [1]:

$$rE \cdot \hat{u}_E = \frac{\sum_{n=0}^{N} G_n^2 \hat{u}_n}{\sum_{n=0}^{N} G_n^2} \tag{2}$$

Where G_n is the gain of the n^{th} loudspeaker and $\hat{u}_n$ is a unitary vector which represents the direction of an incoming wave radiated by the n^{th} loudspeaker. A different approach is used in the In-Phase decoder, which recreates the condition that the loudspeakers feed the signals in phase decreasing the localization artifacts [1]. A detailed description of these decoding methods is beyong to the scope of this paper, but the reader can find a comprehensive discussion in [7].

The implementation of these types of decoder is made by applying a monotonically decreasing weighting function (like a "fade out") to the spherical harmonics coefficients. Consequently, each decoder yields a different sound field reconstruction performance. The concept of this weighting function can be explained using an analogy to the Fourier transform of a Dirac Delta function with different window types. Figure 1 shows delta signals created by applying several different frequency-domain windows. According to the window type selected, the energy of the coefficients is weighted in different proportion leading to an altered signal when the inverse Fourier transform is applied.

[1] Radius of 0.1 m and frequency of 2 kHz require at least 25 loudspeakers

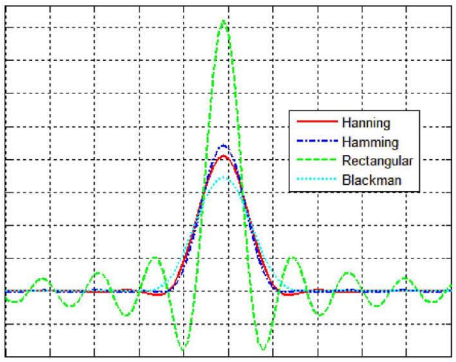


Figure 1: *A Delta Dirac signal after the application of the Fourier transform, truncation using different windows and subsequently inverse transform.*

Extensive work has been made to evaluate the performance of Ambisonics by means of perceptual or physical approaches. These assessments are commonly based on numerical simulations [4, 6, 8] or by listening test [9-11]. However, results obtained from experimental measurements of the acoustic pressure or the acoustic intensity field generated by HOA systems are less frequent in the scientific literature.

This paper evaluates the performance of three different Ambisonics decoding methods (Basic, Max-rE and In-Phase) by analyzing experimental measurements of objective parameters. To that end, a 5th order Ambisonics system was deployed using a spherical array of 40 loudspeakers [12]. The measurements were conducted in the anechoic chamber of the ISVR using a translating microphone array composed by 29 transducers (Ref. Brüel & Kjær 4189-L001) across 40 positions (see figure 2). The total number of measured points corresponded to 1160 with a spatial resolution of 0.05 m leading to an approximate spatial alias frequency of 3.4 kHz.

From the collected data, the sound pressure field and acoustic intensity field were computed and compared with the target field by means of pressure and intensity errors. The rest of this paper is organized as follows: Section 2 presents the methods used for the experiment. Section 3 shows the results of the measurements and the error analysis between the measured and target field. Finally, the conclusions of the current work are presented in Section 4.

Figure 2: *Measurement of the decoding methods.*

2. METHODS

The audio reproduction using an Ambisonics system can be mainly divided into two stages. Firstly, the audio signals are encoded in a finite number of spherical harmonic coefficients. This codification depends on the number of loudspeakers available but not on the size or shape of the array. A 5th order system involves the use of 36 spherical harmonics $(N+1)^2$ to encode the signal. In the second stage, according to the number of the loudspeakers and the shape of the array, the signal is decoded and reproduced. One well established technique to decode the signal is called the mode-matching approach [13]. The reconstruction of a plane wave in direction (θ_i,ϕ_i) using a set of L plane waves each of them with different complex amplitude q_a and direction (θ_a,ϕ_a) can be expressed using the Jacobi-Anger expansion [14] as:

$$4\pi\sum_{n=0}^{\infty}(j^n)j_n(kr)\sum_{m=-n}^{n}Y_n^m(\theta,\phi)Y_n^m(\theta_i,\phi_i)^* = 4\pi\sum_{a=1}^{L}q_a(\omega)\sum_{n=0}^{\infty}(j^n)j_n(kr)\sum_{m=-n}^{n}Y_n^m(\theta,\phi)Y_n^m(\theta_a,\phi_a)^* \quad (3)$$

where k is the wavenumber, ω is the angular frequency, $j_n(kr)$ is the spherical Bessel function of first kind, $j=\sqrt{-1}$ and $Y_n^m(\theta,\phi)$ are the spherical harmonics defined as:

$$Y_n^m(\theta,\phi)=\sqrt{\frac{(2n+1)(n-m)!}{4\pi(n+m)!}}P_n^m\cos(\theta)e^{jm\phi} \quad (4)$$

in which P_n^m is the associated Legendre function. Simplifying equation 3 yields the following matching equation for each n and m:

$$Y_n^m(\theta_i,\phi_i)^*=\sum_{a=1}^{L}q_a(\omega)Y_n^m(\theta_a,\phi_a)^* \quad (5)$$

for $n=0\ldots N$ and $|m|\le n$. This is a finite set of linear equations that can be written in a matrix form as $\boldsymbol{b}_{(m\times L)}=\boldsymbol{Y}_{(m\times L)}\boldsymbol{q}_{(m\times L)}$. In order to have at least one solution, the number of spherical harmonics $(N+1)^2$ is required to be lower than, or equal to, the number of speakers, namely $L\ge(N+1)^2$. Finally, the gains are calculated with the inverse matrix of $\boldsymbol{Y}_{((N+1)^2\times L)}$ if $L=(N+1)^2$ or pseudo-inverse matrix if $L>(N+1)^2$. The stability of the inversion of the matrix Y depends on the loudspeaker array and can be checked by the condition number [15].

The algorithm to test the performance of the Ambisonics decoding methods was developed using the software package Max. Figure 3 shows a diagram of the decoder with its respective modules. The first part corresponds to the encoding stage using up to 5th order of spherical harmonics. Then, the resulting signals were weighted by a G_n function according to the chosen type of decoder (Basic, Max-rE or In-Phase). The values for the G_n functions were calculated using the methodology suggested by Jerome Daniel [7]. Table 1 reports the values of the gains G_n for each type of decoder, according to the order of the spherical harmonics. Finally, at the last stage, the signals are decoded

using the decoding matrix obtained by the mode matching approach and reproduced by the loudspeaker array.

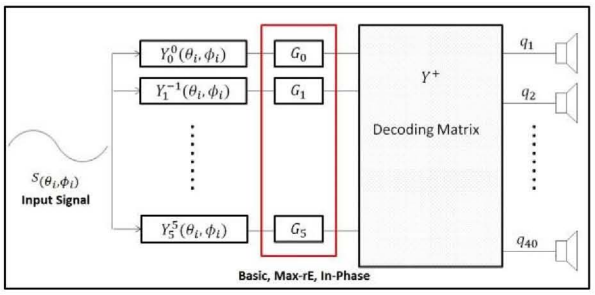


Figure 3: *Sketch of the decoder*

Table 1: *Decoder Gains.*

Order of SH	Basic	Max-rE	In-Phase
1	1	1	1
2	1	0.932	0.75
3	1	0.8029	0.4167
4	1	0.6259	0.1167
5	1	0.5186	0.0455

2.1. Sound pressure and acoustic intensity field

The sound pressure field was directly computed from the measurements. In the case of the acoustic intensity, the values were determined by taking the real part of the product between the sound pressure $p(\mathbf{x})$ and the conjugate of the particle velocity $\mathbf{u}(\mathbf{x})^*$ (see equation 6). The particle velocity was calculated based on the Euler equation (equation 7) by approximating the gradient of the pressure as the difference between neighbouring sound pressure measurement positions (equation 8).

$$\mathbf{I}(\mathbf{x}) = \frac{1}{2}\mathrm{Re}\left\{p(\mathbf{x})\mathbf{u}(\mathbf{x})^*\right\} \tag{6}$$

$$\mathbf{u}(\mathbf{x}) = -\frac{\nabla p(\mathbf{x})}{j\omega\rho_0} \tag{7}$$

$$\mathbf{u}(\mathbf{x}) \approx -\frac{1}{j\omega\rho_0}\frac{\left[p(\mathbf{x}+d\mathbf{x})-p(\mathbf{x})\right]}{d\mathbf{x}} \tag{8}$$

where $\nabla p(\mathbf{x})$ is the gradient of the pressure and ρ_o is the static density of the air.

3. RESULTS

The reconstruction of the acoustic pressure and acoustic intensity flow field for 250 Hz and 2 kHz are presented in Figures 4 and 5, respectively. Red color corresponds to zones of maximum acoustic pressure and blue to the minimum. The black circle represents the region of validity calculated from equation 1. In case of 250 Hz, the radius of validity is bigger than the dimension of the array so it is expected to have an accurate reconstruction over the whole measured area. Figures of 1 kHz are also presented in Appendix 1.

The measurement procedure involved the recording of the sound field generated by each type of decoder using the microphone array. The excitation signal corresponded to a virtual point source (white noise) located at 45° in azimuth [0°, 360°], 0° in elevation [90°, -90°] and 1.8 m far away. A comparison between figures 4 and 5 clearly identifies the limitation of Ambisonics to reproduce high frequencies. The radius of validity 'r' provides an insight on the area where the sound field reconstruction is accurate. However, it was found that this assumption is not always valid and depends strongly on the decoder. A more robust analysis of the data is performed in the next subsection.

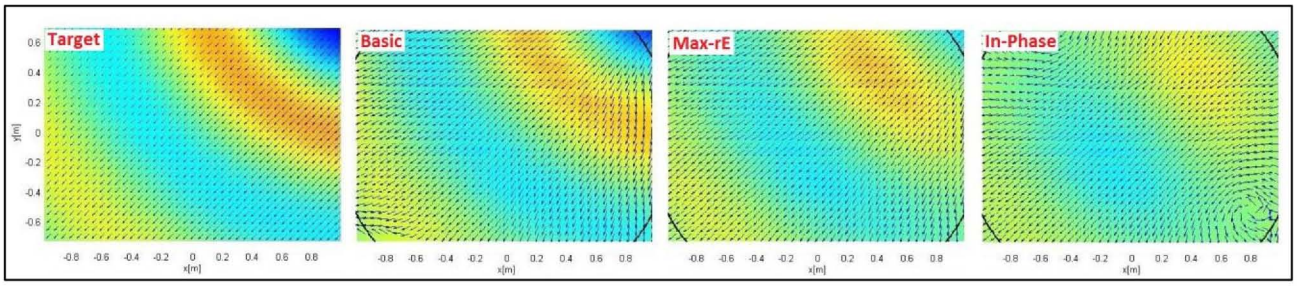


Figure 4: *Sound pressure and acoustic intensity field flow reconstruction for 250 Hz.*

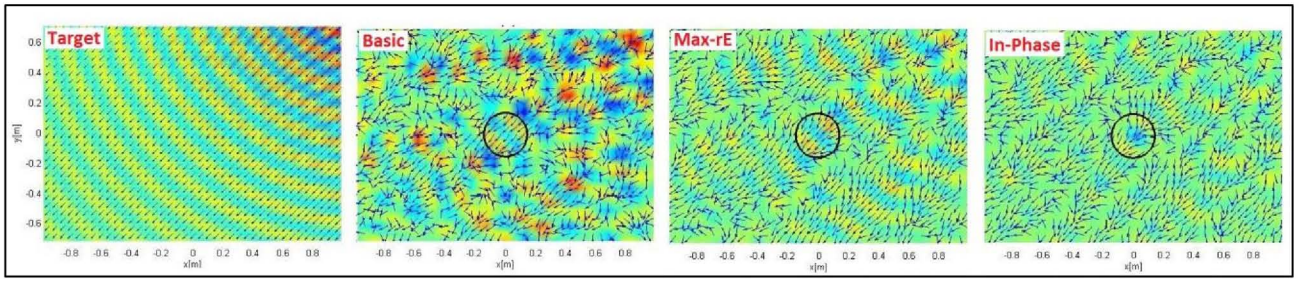


Figure 5: *Sound pressure and acoustic intensity flow field reconstruction for 2 kHz*

3.1. Error analysis

An error analysis was conducted on the sound pressure and the acoustic intensity data. The following error metrics have been adopted to assess the performance of the decoders:

Sound pressure errors:
Amplitude error:

$$E_{pa}(\mathbf{x}) = 10\log_{10}\left(\frac{|p_m(\mathbf{x})|}{|p_t(\mathbf{x})|}\right) \quad (9)$$

Phase error:

$$E_{pp}(\mathbf{x}) = \text{angle}\left(\frac{p_t(\mathbf{x})p_m(\mathbf{x})^*}{|p_t(\mathbf{x})||p_m(\mathbf{x})|}\right) \quad (10)$$

where $p_m(\mathbf{x})$ is the measured pressure, $p_t(\mathbf{x})$ is the target pressure and $p_m(\mathbf{x})^*$ indicates the conjugate of the measured pressure.

Acoustic intensity error:
Angular error [16]:

$$E_{lan} = \frac{\left(I_{x\_tar}I_{x\_mea}\right)+\left(I_{y\_tar}I_{y\_mea}\right)}{\left(\sqrt{\left(I_{x\_tar}\right)^2+\left(I_{y\_tar}\right)^2}\right)\left(\sqrt{\left(I_{x\_mea}\right)^2+\left(I_{y\_mea}\right)^2}\right)} \quad (11)$$

in which I_{x_mea} and I_{y_mea} are the components of the measured acoustic intensity in $\hat{x}$ and $\hat{y}$ directions respectively. I_{x_tar} and I_{y_tar} are the components of the target acoustic intensity.

Figures 6 and 7 show the amplitude error of the sound pressure in dB. At 250 Hz, excellent agreement between the target field and the synthesized field is found for the Basic decoder. For the Max-rE and In-Phase decoders, the reconstructions are accurate at the center of the listener area, but over a region with a smaller radius than the predicted by the equation 1. At 2 kHz, the Basic decoder does not reconstruct the sound field as is expected, even within the radius of validity. The In-phase decoder presents a better performance compared to the Basic decoder, but the Max-rE decoder offers the best performance at this frequency.

The sound pressure phase error is illustrated in figures 8 and 9. The unit of the color bar corresponds to radians (from -π to π). At 250 Hz, the Basic decoder yields to the most accurate phase reconstruction. The Max-rE decoder also provides a good performance in terms of phase error except for the top left corner of the measured area, where a small mismatch can be observed. As in the case of the sound pressure amplitude, the In-Phase decoder leads to the largest errors at this frequency. At 2 kHz, the synthesized phase for all decoding methods tends to be more consistent with the measured data within the radius of validity. However, the Max-rE decoder achieves the best match for this case.

Regarding acoustic intensity, figures 10 and 11 show the angular error at 250 Hz and 2 kHz, respectively. The color bar represents the difference in degrees between 0° and 180°. At 250 Hz, the reference radius matches with the intensity flow created by the Basic and Max-rE decoders. This is not the case for the In-Phase decoder. Max-rE yields the best results for the intensity flow, but not in terms of the amplitude of the intensity where the Basic decoder is better. At 2 kHz, the angular error is almost zero inside of the reference radius for the Basic and Max-rE decoders. Nevertheless, using Max-rE, in some zones outside of this radius the intensity flow and the amplitude errors are comparatively small.

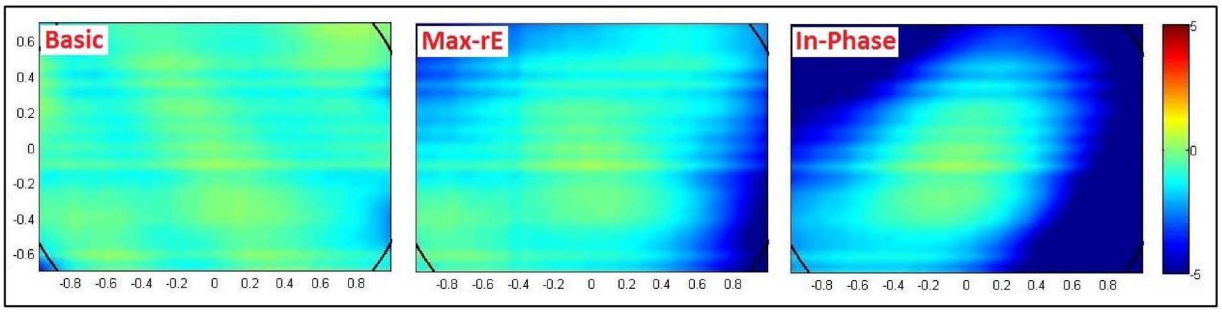


Figure 6: *Sound pressure-amplitude error for 250 Hz*

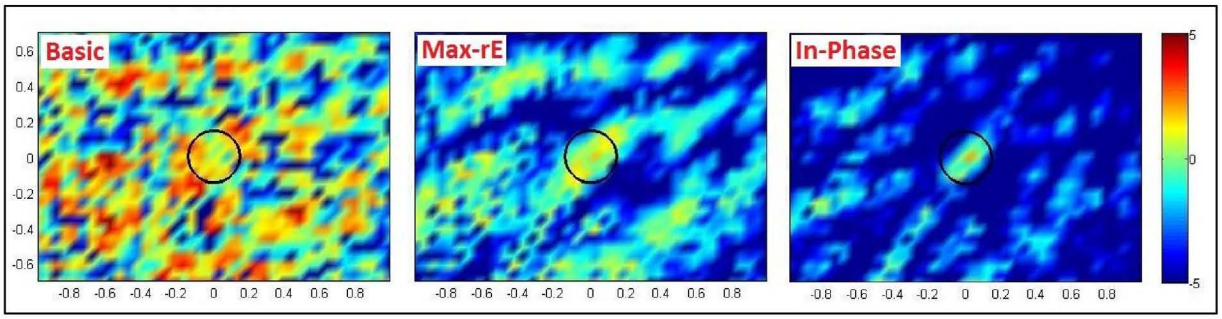


Figure 7: *Sound pressure-amplitude error for 2 kHz*

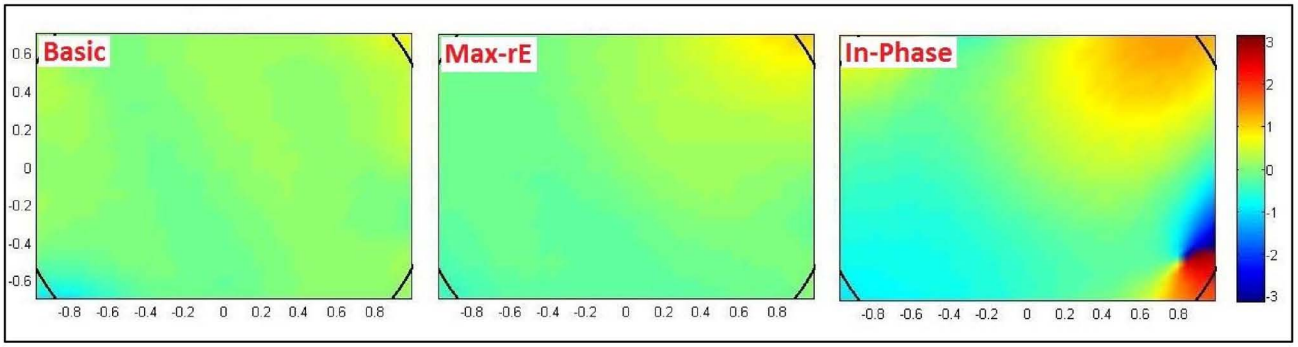


Figure 8: *Sound pressure-phase error for 250 Hz*

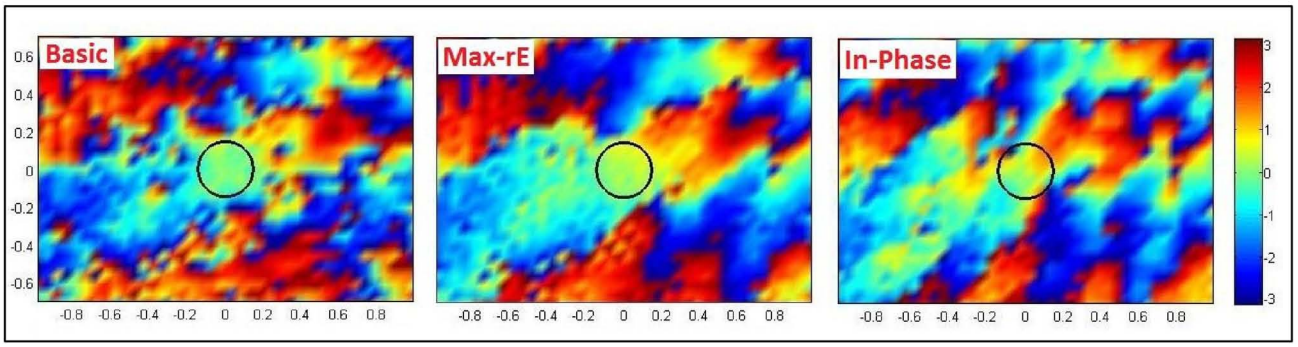


Figure 9: *Sound pressure-phase error for 2 kHz*

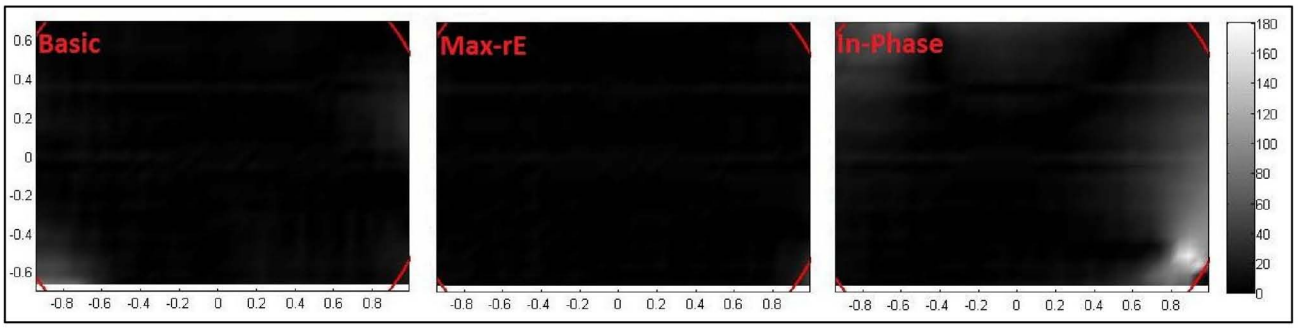


Figure 10: *Acoustic intensity-angular error for 250 Hz*

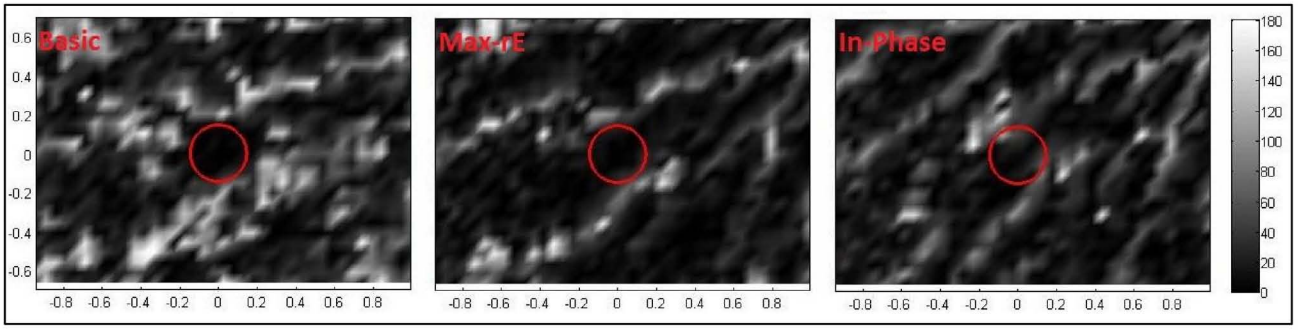


Figure 11: *Acoustic intensity-angular error for 2 kHz*

3.2. Discussion of the results

From the error analysis, it is possible to identify that the performance of the decoders is highly dependant on frequency. At low frequencies, the basic decoder provides the best performance taking into consideration the sound pressure errors. Also, good agreement between the radius of validity and the area where the reconstruction is accurate has been described. However, the results of Max-rE and In-Phase decoders do not follow the rule of thumb $L \geq (kr+1)^2$ generating a smaller area where the pressure and intensity error are low. For this array, the In-Phase decoder is the worst method for reconstructing the low frequency sound field.

At high frequencies, the performance of the Basic decoder decreases noticeably compared to the other decoding methods. It shows the largest errors on both pressure and intensity amplitude[2] compared with the target fields. Nevertheless, the phase pressure and the angular intensity errors are low within the radius of validity. A comparison with the other decoding methods indicates that Max-rE offers the best results. This can be explained by the optimization of the energy vector which is the goal of this decoder. The errors on both pressure and intensity are the lowest when compared to other decoders.

Based on these results, if the aim is to reproduce an audio signal composed by a wide range of frequencies, the use of multiple decoding methods according to the frequencies may be advisable. To that end, the signal can be filtered and processed by different decoders based on the best performance in this specific frequency range. Examples of frequency dependent decoders can be found in [1,17].

4. CONCLUSIONS

The performance of three different Ambisonics decoding methods was evaluated in the light of experimental results. The findings confirm that the accuracy of sound field reproduction by a specific decoders depends on the frequency components. For this array, at low frequencies, the Basic decoder provides the best performance in terms of sound field reconstruction. In contrast, Max-rE presents the best performance at high frequencies. The implementation of combined Ambisonics decoding methods to reproduce a wide-frequency audio signal seems to be the most suitable option.

The concept of region of validity gives an indication of the area where the reconstruction is accurate. However, this assumption is not always valid in practice and significantly depends on the frequency and the type of decoder. The best match between the rule of thumb $L \geq (kr+1)^2$ and the reconstructed sound field was achieved, as expected, with the Basic decoder.

As the sound pressure, the acoustic intensity is another useful parameter that can be used to evaluate the performance of Ambisonics systems. Especially important is the angular error of the intensity, which cannot be evaluated for the acoustic pressure field as this does not contain directional information. An analysis in terms of pressure and intensity allows a more robust examination of reconstructed sound fields.

Finally, it is relevant to emphasize that the measurements were carried out in an anechoic environment using a spherical loudspeaker array which is far from the usual reproduction conditions. The performance of the decoders in regular rooms with comparatively low reverberation using a non-regular array is a topic for future research. Also, a near field compensation may be implemented in order to optimize the sound field for sources close to the listener.

[2] No information about the error in the amplitude of the acoustic intensity was reported in this paper. However, it was calculated to analyse the performance of the decoding methods.

5. REFERENCES

[1] D. Jerome, J. Rault and J. Polack "Ambisonics Encoding of Other Audio Formats for Multiple Listening Conditions," *in 105th Convention of Audio Engineering Society*, San Francisco, September 1998.

[2] M. Gerzon, "Periphony: With-Height Sound Reproduction ," *J. Audio Engineering Society*, vol. 21, no. 1, pp.2-10, February 1973.

[3] S. Favrot and j. Buchholtz, "LoRA: A Loudspeaker-Based Room Auralization System," *Acta Acustica United with Acustica*, vol. 96, pp.364-375, 2010.

[4] D. Jerome, R. Nicol and S. Moreau, "Further Investigations of High Order Ambisonics and Wavefield Synthesis for Holophonic Sound Imaging," *in 114th Convention of Audio Engineering Society*, Amsterdam, March 2003.

[5] E. Williams. *Fourier Acoustics*. London: Academic Press, 1999.

[6] D. Ward and T. Abhayapala, "Reproduction of a plane-wave sound field using an array of loudspeakers," *IEEE Transactions on Speech and Audio Processing*, vol. 9, pp. 697-707, Sep 2001.

[7] D. Jerome, Acoustic field representation, application to the transmission and the reproduction of complex sound environments in a multimedia context. Ph.D. dissertation, University of Paris, Paris, France, 2011.

[8] A. Solvang, "Spectral Impairment for Two-Dimensional Higher Order Ambisonics," *J. Audio Engineering Society*, vol. 56, no. 4, pp.267-279, April 2008.

[9] A. Horsburgh, R. Davis, M. Moffat and D. Fraser, "Subjective Assessments of Higher Order Ambisonic Sound Systems in Varying Acoustical Conditions," *in 133rd Convention of Audio Engineering Society*, San Francisco, October, 2012.

[10] G. Kearney, E. Bates, F. Boland and D. Furlong, "Comparative Study of the Performance of Spatialization Techniques for a Distributed Audience in a Concert Hall Environment, "

in 31st International Conference of Audio Engineering Society, London, June, 2007.

[11] S.Bernet, J. Daniel, E. Parizet , L. Gros and O. Warusfel, "Investigation of the Perceived Spatial Resolution of Higher Order Ambisonic Sound Fields: a Subjective Evaluation Involving Virtual and Real 3D Microphones, " *in 30st International Conference of Audio Engineering Society*, Saariselka, March, 2007.

[12] F Fazi, Sound Field Reproduction. Ph.D. dissertation, University of Southampton, Southampton, United Kingdom, 2010.

[13] M. Poletti, "Three-Dimensional Surround Sound systems Based on Spherical Harmonics ," *J. Audio Engineering Society*, vol. 53, no. 11, pp.1004-1025, November 2005.

[14] B. Rafaely, "Plane Wave Decomposition of the Sound Field on a Sphere by Spherical Convolution", Institute of Sound and Vibration Research, ISVR, United Kingdom, Tech. Mem. 910, 2003.

[15] B. Stofringsdal and P.Svensson, "Conversion of Discretely Sampled Sound Field Data to Auralization Formats," *J. Audio Engineering Society*, vol. 54, no. 5, pp.380-400, May 2008.

[16] M. Shin, F. Fazi, P. Nelson and J. Seo "Control of Velocity for Sound Field Reproduction *," in 52nd InterCational conference of Audio Engineering Society*, Guildford, September, 2013.

[17] M. Gerzon, "Psychoacoustic Decoders for Multispeaker Stereo and Surround Sound," ," in 93rd Convention of Audio Engineering Society, San Francisco, October, 1992.

APPENDIX 1

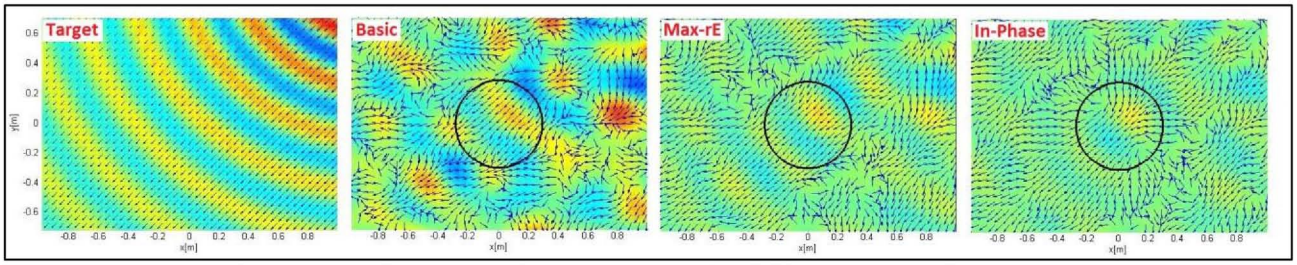


Figure 12: *Sound pressure and acoustic intensity flow field reconstruction for 1 kHz.*

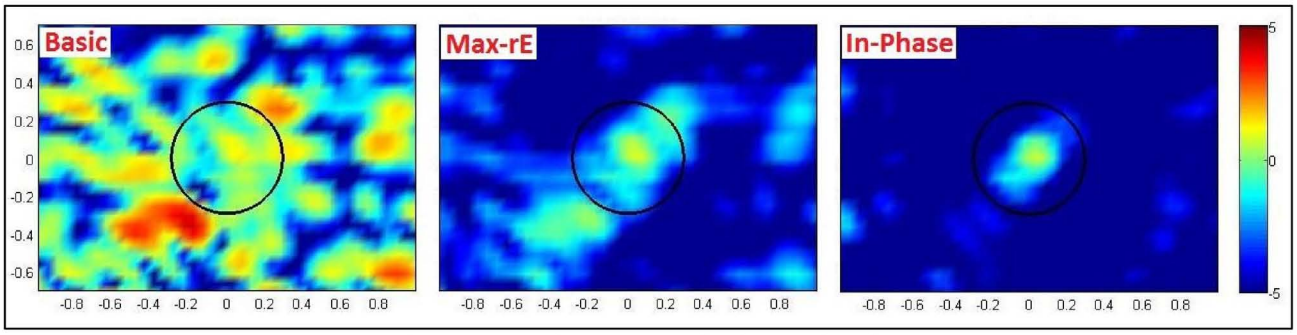


Figure 13: *Sound pressure-amplitude error for 1 kHz*

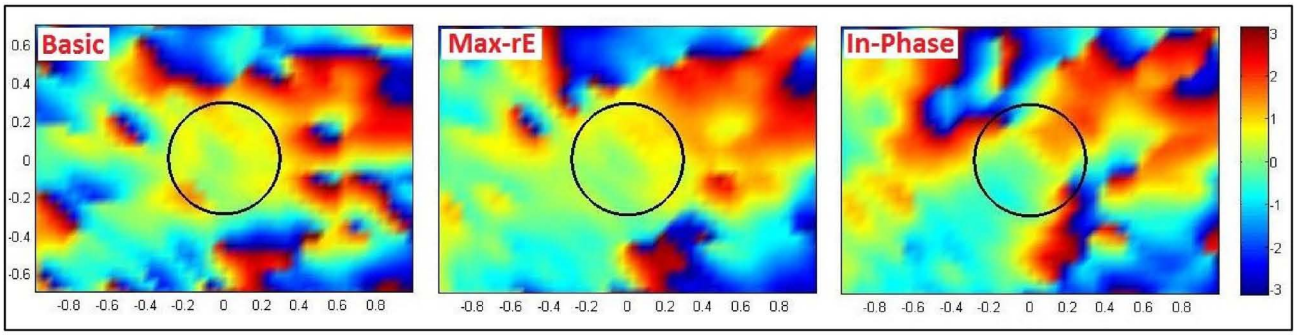


Figure 14: *Sound pressure-phase error for 1 kHz*

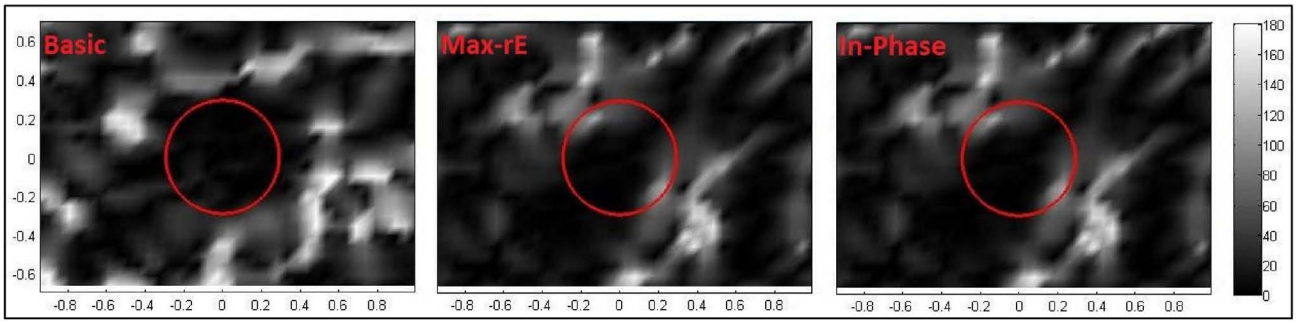


Figure 15: *Acoustic intensity-angular error for 1 kHz*

LOCALIZATION USING DIFFERENT AMPLITUDE-PANNING METHODS IN THE FRONTAL HORIZONTAL PLANE

Matthias Frank

Institute of Electronic Music and Acoustics,
University of Music and Performing Arts Graz
Graz, Austria
frank@iem.at

ABSTRACT

Amplitude panning is the simplest method to create phantom sources in the horizontal plane. The most commonly employed amplitude-panning methods are Vector-Base Amplitude Panning (VBAP), Multiple-Direction Amplitude Panning (MDAP), and Ambisonics. This article investigates the localization of frontal phantom sources created by VBAP, MDAP, and Ambisonics (with and without max-r_E weighting) at the central listening position in a listening experiment. The experiment was conducted under typical non-anechoic studio listening conditions and utilized pink noise and a regular array of 8 loudspeakers for all methods. The experimental results are compared to different predictors: a binaural localization model using measured binaural room impulse responses, the direction of the measured sound intensity vector, and the directions of the simpler velocity and energy vectors. The article hereby addresses the questions of how close the actually localized directions of the different panning methods are compared to the desired directions, and how good the predictors match the experimental results.

1. INTRODUCTION

Amplitude-panning methods use simple level differences between the loudspeakers to evoke auditory objects between the loudspeakers, so-called phantom sources [1]. Although the computational effort is similar, the methods differ in their theoretical basis and the number of loudspeakers they use for each phantom source. This contribution examines perceptual differences between the most commonly used amplitude-panning methods, in particular the phantom source localization. This is done by employing existing experimental results from the thesis of the author [2] that used vector-base amplitude panning, multiple-direction amplitude panning, and Ambisonics with different weightings on the same loudspeaker arrangement. Although there are some studies about the localization of Ambisonics [3, 4, 5], comparisons to the other panning methods are rare [6]. The experiment focuses on frontal directions (with a maximum displacement from the median plane of $45°$) in order to compare the panning methods within the angular range where human sound source localization works most accurately [7].

The experimental results are compared to the directions of the simple velocity and energy vector. Although the suitability of these measures for the prediction of phantom source localization has not been proven yet, they are often applied in practice for Ambisonics decoder design [8]. This contribution examines their suitability and compares them to a state-of-the-art binaural localization model and measurements of the sound intensity vector.

The second section introduces the employed amplitude-panning methods and their theoretical basis. The experimental localization results for the panning methods are presented in the third section. In order to examine the controllability of the phantom source location, subjective variation is discussed and the results are compared to the desired panning direction. Section four presents localization predictors that are based on dummy head or microphone array measurements at the listening position within the actual sound field, as well as simpler predictors incorporating solely the loudspeaker gains and positions. Finally, their predictions are compared to the experimental results.

Throughout this contribution, the directions of L loudspeakers, as well as the panning directions are expressed as unit vectors $\boldsymbol{\theta} = [\cos(\phi), \sin(\phi)]^{\mathrm{T}}$ depending on the azimuth angle ϕ in the x-y plane, cf. Figure 1. The scalar weight g_l of each loudspeaker $l \in \{1...\mathrm{L}\}$ denotes its adjustable gain.

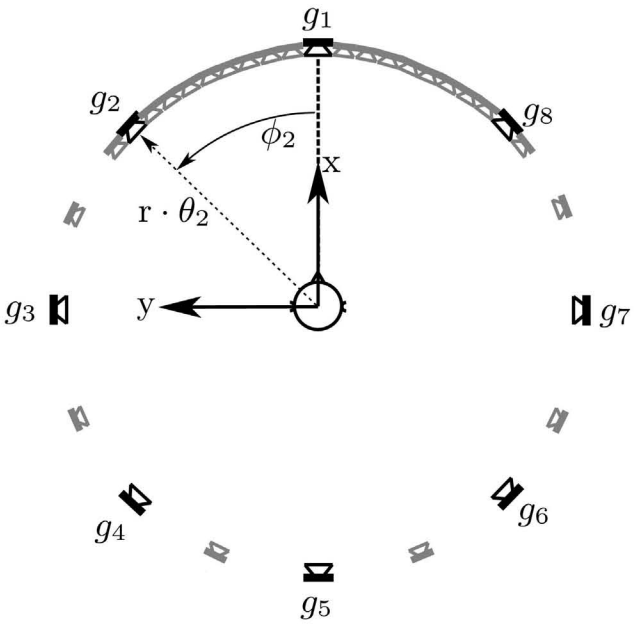


Figure 1: *Experimental setup in the reference coordinate system. The smaller, gray loudspeakers are visible but inactive.*

2. HORIZONTAL AMPLITUDE-PANNING METHODS

All presented amplitude-panning methods are also applicable to three-dimensional loudspeaker arrangements. However, this overview focuses on the two-dimensional case that is employed in this contribution.

2.1. Vector-Base Amplitude Panning (VBAP)

Vector-Base Amplitude Panning (VBAP) [9] can be seen as the generalization of the tangent law [10] for amplitude panning in two-channel stereophony [11]. The tangent law is based on a simple geometrical head model and is still the most popular panning law for pairwise panning. In order to create a phantom source within a loudspeaker pair located at $\mathbf{L}_{\mathrm{ij}} = [\boldsymbol{\theta}_{\mathrm{i}},\, \boldsymbol{\theta}_{\mathrm{j}}]$, VBAP calculates the weights $\boldsymbol{g}_{\mathrm{ij}} = [g_{\mathrm{i}},\, g_{\mathrm{j}}]^{\mathrm{T}}$ depending on the panning direction $\boldsymbol{\theta}_s$:

$$\boldsymbol{g}_{\mathrm{ij}} = \mathbf{L}_{\mathrm{ij}}^{-1}\, \boldsymbol{\theta}_s. \tag{1}$$

Typically, a subsequent normalization of the gains is necessary to keep the overall energy constant. The aperture angle of the loudspeaker pair should not exceed 90° [12] resulting in non-negative weights. To extend the panning range around one pair, more loudspeaker pairs can be attached [9].

Basically, the number of active loudspeakers is depending on the panning direction: two loudspeakers are active for directions between two loudspeakers and one is active for directions coinciding with a loudspeaker.

2.2. Multiple-Direction Amplitude Panning (MDAP)

For a more uniform panning, those cases of VBAP with only one active loudspeaker can be avoided. This is done by extending VBAP to Multiple-Direction Amplitude Panning (MDAP) [13].

MDAP superimposes the results of VBAP for B panning directions uniformly distributed around the desired panning direction $\boldsymbol{\theta}_s$ within a spread of $\pm\phi_{\mathrm{MDAP}}$. Typically, the maximum spread is related to the loudspeaker spacing $\Delta\phi_L$ of a uniform loudspeaker arrangement. In this contribution, MDAP is always used with $\mathrm{B} = 10$ panning directions uniformly distributed within a spread of $\phi_{\mathrm{MDAP}} = {}^1\!/_2\Delta\phi_L$.

For this setting, three loudspeakers are active for most panning directions, even for directions on a loudspeaker. If the panning direction lies exactly in the middle between two loudspeakers, only these two loudspeakers are active. In these special cases, MDAP yields the same loudspeaker gains as VBAP.

2.3. Ambisonics

Ambisonics [14, 15, 16, 17] is a recording and reproduction method which is based on the representation of the sound field excitation as a superposition of orthogonal basis functions. In the horizontal case, these functions are the periodic trigonometric basis of the Fourier series, the so-called circular harmonics. Their maximum order N determines the spatial resolution and the number $2\mathrm{N}+1$ of signals and minimum required loudspeakers.

For one source at a direction $\boldsymbol{\theta}_s$, the Ambisonic spectrum $\boldsymbol{y}_{\mathrm{N}}(\boldsymbol{\theta}_s)$ is calculated by evaluating the circular harmonics at $\boldsymbol{\theta}_s$ This calculation is frequency-independent and assumes that all sources and the loudspeakers lie on a circle of the same radius r.

The decoder derives the gains $\boldsymbol{g} = \{g_1, ...g_{\mathrm{L}}\}$ for the L loudspeakers of an arrangement from the Ambisonic spectrum $\boldsymbol{y}_{\mathrm{N}}(\boldsymbol{\theta}_s)$ by multiplication with the decoder matrix $\mathbf{D}$:

$$\boldsymbol{g} = \mathbf{D}\, \mathrm{diag}\{\boldsymbol{a}_{\mathrm{N}}\}\, \boldsymbol{y}_{\mathrm{N}}(\boldsymbol{\theta}_s). \tag{2}$$

The matrix is derived from the circular harmonic spectra $\boldsymbol{y}_{\mathrm{N}}(\theta_l)$ of each loudspeaker $\boldsymbol{Y}_{\mathrm{N}} = [\boldsymbol{y}_{\mathrm{N}}(\theta_1),\, \boldsymbol{y}_{\mathrm{N}}(\theta_2),\, \ldots,\, \boldsymbol{y}_{\mathrm{N}}(\theta_{\mathrm{L}})]$. It can be calculated by transposition or inversion of $\boldsymbol{Y}_{\mathrm{N}}$, resulting in a sampling or mode-matching decoder [18], respectively. The energy-preserving decoder [19] uses more sophisticated techniques, such as singular value decomposition. See the appendix of [20] for an overview about different decoders. In this contribution, the regular arrangement of L loudspeakers (cf. Figure 1) ensures that the simple sampling decoder is also mode-matching and energy-preserving for all orders $\mathrm{N} \leq (\mathrm{L}-1)/2$.

In order to control the main and side lobes emerging from the truncation of the circular harmonics, a weighting vector $\boldsymbol{a}_{\mathrm{N}}$ is applied in the harmonics domain [17]. The basic weighting uses a vector of ones $\boldsymbol{a}_{\mathrm{N}} = \mathbf{1}$, whereas the max-$\mathrm{r_E}$ weighting suppresses the side lobes at the cost of a wider main lobe by attenuating higher orders. This is done by an order-depend weight $a(n) = \cos(\frac{n\pi}{2\mathrm{N}+2})$. Another weighting, called in-phase, yielded no convincing results in previous experiments [5] and is therefore not used here.

Basically, Ambisonics always uses all available loudspeakers for the creation of a single phantom source. However, the equiangular arrangement of L even-numbered loudspeakers yields an exception when using max-$\mathrm{r_E}$ Ambisonics with an order of $\mathrm{N} = \mathrm{L}/2-1$: for panning directions exactly in the middle between two neighboring loudspeakers, only these two loudspeakers are active. In these cases, max-$\mathrm{r_E}$ Ambisonics yields the same loudspeaker gains as VBAP and MDAP.

3. EXPERIMENT

The listening experiment evaluates the localization of phantom sources created by VBAP, MDAP, basic Ambisonics, and max-$\mathrm{r_E}$ Ambisonics at the central listening position.

3.1. Setup and Conditions

All panning methods employ a regular ring of 8 Genelec 8020 loudspeakers at a radius of $\mathrm{r} = 2.5\,\mathrm{m}$. Figure 1 shows the experimental setup with additional inactive but visible loudspeakers placed in 5° steps in and around the angular range of the target directions. The height of all loudspeakers was set to $1.2\,\mathrm{m}$ which was also the ear height of the subjects. The experiment was performed in the IEM CUBE, a $10.3\,\mathrm{m} \times 12\,\mathrm{m} \times 4.8\,\mathrm{m}$ large room with a mean reverberation time of $470\,\mathrm{ms}$ that fulfills the recommendation for surround reproduction in ITU-R BS.1116-1 [2, 21]. The central listening position lies within the effective critical distance.

The control of the entire experiment and the creation of the loudspeaker signals used the open source software pure data[1] on a standard PC with RME audio interface and D/A converters. The perceived direction was assessed by a pointing method using a toy-gun that was captured by an infrared tracking system. Details about the pointing method can be found in [22].

[1] freely available on `http://puredata.info/downloads`

Both Ambisonics variants use a maximum order of 3 and MDAP is applied with $\mathrm{B} = 10$ panning directions uniformly distributed within a spread of $\phi_{\mathrm{MDAP}} = 22.5°$. The effect of the Ambisonics order has already been studied in [5] and is not part of this contribution. All panning methods were evaluated for 9 directions (with an even spacing of $5.625°$) between $0°$ and $-45°$ (to the right). Each of the 36 = 9 (directions) $\times$ 4 (panning methods) conditions was evaluated twice by each subject in random order. The stimulus consisted of 3 pink noise bursts, each with $100\,\mathrm{ms}$ fade-in, $200\,\mathrm{ms}$ at $65\,\mathrm{dB(A)}$, $100\,\mathrm{ms}$ fade-out, and $200\,\mathrm{ms}$ silence before the next fade-in. The stimulus playback could be repeated at will by the subjects.

There were 14 subjects participating in the experiment. All of them were part of a trained expert listening panel [23, 24].

3.2. Results

An analysis of variance (ANOVA) showed that the repetition was not a significant factor ($p = 0.522$ for VBAP, $p = 0.465$ for MDAP, $p = 0.085$ for basic Ambisonics, and $p = 0.91$ for max-$\mathrm{r_E}$ Ambisonics). This confirms a high intra-rater reliability and thus no subjects were excluded from the results. On the other hand, the subjects were a highly significant factor ($p \ll 0.001$) for all tested panning methods. This agrees with the inter-subjective localization found for lateral [2] and vertical phantom sources [25, 26]. Nevertheless, the following localization curves summarize all 28 answers from all subjects and repetitions for each condition.

Figure 2 shows the median values and the corresponding 95% confidence intervals for the 9 different panning angles using VBAP at the central listening position. Obviously, the panning angle is a significant factor ($p \ll 0.001$). All neighboring conditions were perceived from significantly different directions ($p \ll 0.001$), except for the angles $-45°$ and $-39.375°$ ($p = 0.053$). In comparison to the ideal localization curve (perceived angle = panning angle), the perceived angles tend towards the loudspeakers. This tendency is known from [27] and is even more distinct for lateral directions.

Using MDAP, the panning angle is still a significant factor ($p \ll 0.001$). This holds true for the direct comparison of neighboring panning angles. Compared to VBAP, the median perceived angles for MDAP are closer to the ideal panning curve and yield a reduced tendency towards the loudspeakers, cf. Figure 3.

The experimental results for basic Ambisonics show a stretched trend, i.e. a steeper slope than the ideal curve, cf. Figure 4. The discriminability of the panning angles is comparable to MDAP.

The significant discriminability of the panning angles holds true for max-$\mathrm{r_E}$ Ambisonics. Figure 5 shows that the median experimental results are very close to the ideal localization curve for this panning method.

Table 1 compares the median deviation of the different panning methods from the ideal localization curve, i.e., how much the perceived angle deviated from the panning angle. The angles deviate most for VBAP, in fact more than two times as much as for max-$\mathrm{r_E}$ Ambisonics. The angular match is best for max-$\mathrm{r_E}$ Ambisonics, followed by MDAP, and basic Ambisonics.

Table 1: *Average absolute deviation of median experimental results from ideal localization curve for different panning methods.*

VBAP	MDAP	basic	max-$\mathrm{r_E}$
$2.35°$	$1.28°$	$1.58°$	$1.05°$

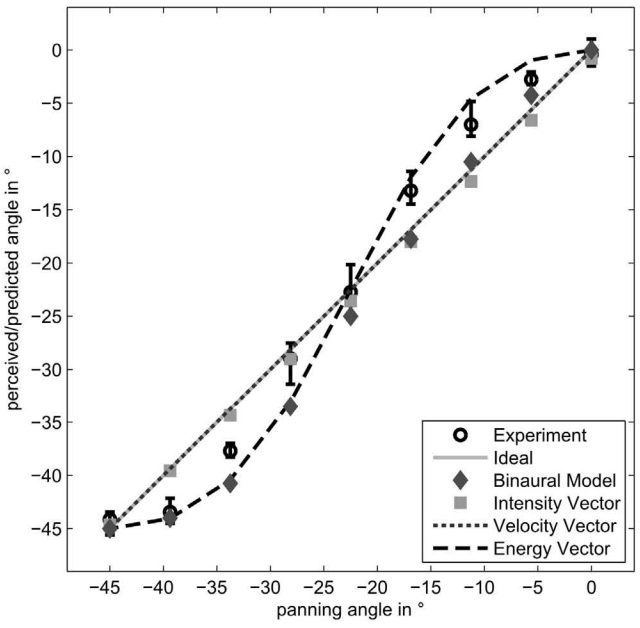


Figure 2: *Localization curves for VBAP: experimental results (median values and 95% confidence intervals), ideal curve (perceived/predicted angle = panning angle), and predictions by binaural model, intensity vector, velocity vector, and energy vector.*

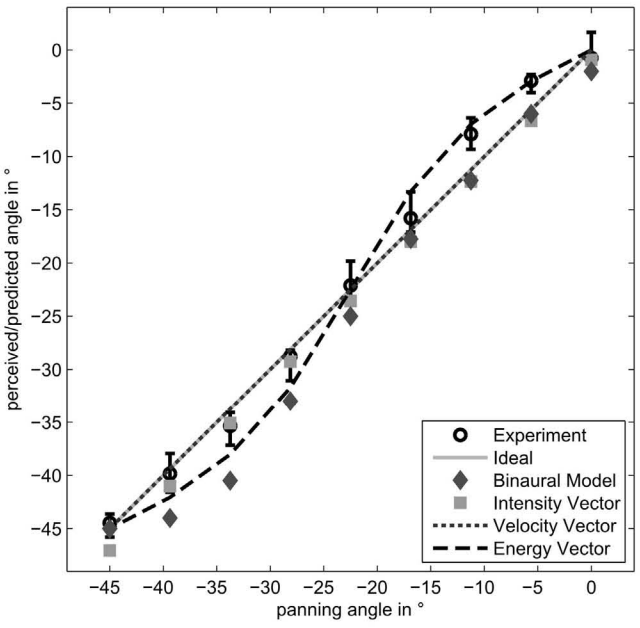


Figure 3: *Localization curves for MDAP: experimental results (median values and 95% confidence intervals), ideal curve (perceived/predicted angle = panning angle), and predictions by binaural model, intensity vector, velocity vector, and energy vector.*

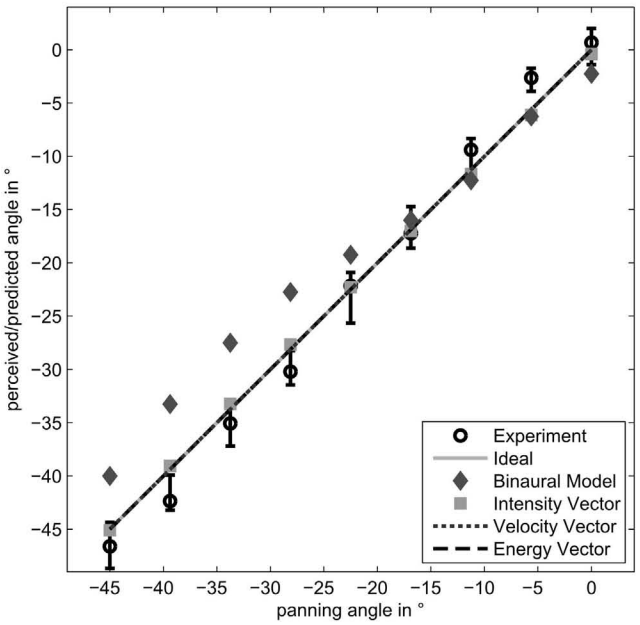


Figure 4: *Localization curves for basic Ambisonics: experimental results (median values and 95% confidence intervals), ideal curve (perceived/predicted angle = panning angle), and predictions by binaural model, intensity vector, velocity vector, and energy vector.*

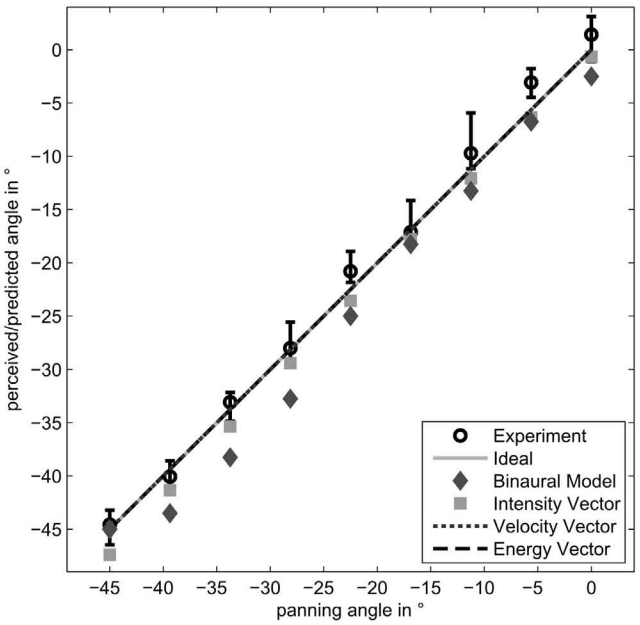


Figure 5: *Localization curves for max-r_E Ambisonics: experimental results (median values and 95% confidence intervals), ideal curve (perceived/predicted angle = panning angle), and predictions by binaural model, intensity vector, velocity vector, and energy vector.*

Despite this ranking, VBAP yields the smallest standard deviations in the experimental results, cf. Table 2. This is mainly the case for panning angles close to the loudspeakers, which provide narrow and accurate localization in VBAP in comparison to phantom sources created by the other methods. Obviously, for higher number of active loudspeakers, the standard deviation increases. This holds true for the total standard deviation, as well as for the inter-subjective and the intra-subjective standard deviation. However, the inter-subjective standard deviation is greater than its intra-subjective counterpart for all panning methods, agreeing with the results from the ANOVA that showed the subjects to be a significant factor, but not the repetition.

Table 2: *Mean total, inter-subjective, and intra-subjective standard deviations of experimental results for different panning methods.*

	VBAP	MDAP	basic	max-r_E
total	2.93°	3.32°	4.61°	4.07°
inter-subj.	2.52°	2.94°	3.94°	3.54°
intra-subj.	1.87°	1.90°	2.59°	2.28°

4. PREDICTIONS

In order to save experiments in the future, it is desirable to find suitable predictors for the localization of phantom sources. This section presents a selection of predictors that differ in the measurement effort and it compares their predictions to the experimental results.

4.1. Localization Predictors

Binaural Model

This contribution employs a binaural localization model after Lindemann [28, 29] which is part of the Auditory Modeling Toolbox[2]. It divides the binaural input signals into 36 frequency bands with a spacing of 1ERB (equivalent rectangular bandwidth) [30]. The auditory nerve is modeled by a half-wave rectifier and a low-pass filter at 800Hz. In each band, the inter-aural level-difference (ILD) is considered by monaural detectors and contra-lateral inhibition. The inter-aural time-difference (ITD) is then computed as the centroid of the inter-aural cross correlation function [31], which delivers one ITD value for each frequency band.

Within each frequency band, the ITD value of the phantom source is compared to the values of a single sound source in a lookup table. The best matching ITD is selected and the corresponding angle is regarded as the angle of the phantom source for the present frequency band. A single angle as prediction result is achieved by the median value of the angles for all frequency bands. The best fit to the median experimental results has been achieved when using 21 frequency bands covering the range from $164\,\mathrm{Hz} - 3558\,\mathrm{Hz}$. On the one hand, the lower frequency limit seems reasonable as very low frequencies do not yield inter-aural differences because of the large wavelengths in comparison to the head diameter. On the other hand, the upper frequency limit underlines the dominant role of low-frequency ITDs, albeit it also supports the importance of ITDs at higher frequencies in comparison to the classical duplex theory [32].

The model is fed with binaural room impulse responses recorded at the central listening position of the experimental setup with a

[2]freely available on `amtoolbox.sourceforge.net/`

B&K 4128C dummy head. It uses the first 80 ms of the impulse responses. As the model cannot distinguish between front and back, the ITD values in the lookup table were limited to the directions between $\pm 45^\circ$, where the conditions of the experiment lie.

Intensity Vector
Sound intensity is a physical measure of the directional sound power flow and can thus be used to determine the direction where sound is coming from [33]. The intensity $\boldsymbol{I}$ is computed from the scalar sound pressure p and the vectorial particle velocity $\boldsymbol{v}$ as $\boldsymbol{I} = p\boldsymbol{v}$ [34]. The sound pressure can be measured by an omni-directional microphone. The particle velocity is typically not measured directly but by the pressure gradients v_x and v_y using figure-of-eight microphones, each one aligned to the axis of the coordinate system. Here p, v_x, and v_y are computed as the convolution of A-weighted pink noise with impulse responses measured with two Schoeps CCM 8 figure-of-eight microphones and an NTI MM2210 omni-directional microphone at the central listening position.

As a predictor for sound source directions, it is not suitable to compute the instantaneous direction of the intensity vector for each sample separately, i.e. 44100 times a second, but rather as a temporal average within a certain time window. The time window was set to 80 ms (S = 3528 samples), which corresponds to the binaural localization model. The components $\overline{I_x}$ and $\overline{I_y}$ of the temporally averaged intensity vector $\overline{\boldsymbol{I}} = [\overline{I_x}\ \overline{I_y}]$ are computed as

$$\overline{I_x} = \sum_{s=1}^{\mathrm{S}} p(s)v_x(s) \qquad \text{and} \qquad \overline{I_y} = \sum_{s=1}^{\mathrm{S}} p(s)v_y(s). \tag{3}$$

The direction of the intensity vector is calculated as $\arctan(\overline{I_y}, \overline{I_x})$ and is equal to the direction of the velocity vector under free-field conditions.

Velocity Vector
The direction of the velocity vector was proposed as a simple predictor for the localization of low frequencies (≤ 700 Hz) [35, 36]. It is calculated as linear summation of the weighted loudspeaker directions:

$$\boldsymbol{r}_{\mathrm{V}} = \frac{\sum_{l=1}^{\mathrm{L}} g_l \boldsymbol{\theta_l}}{\sum_{l=1}^{\mathrm{L}} g_l}. \tag{4}$$

As it is solely based on the loudspeaker directions and gains, it does not require any acoustical measurements. It assumes free-field conditions or at least a dominant direct sound. For two loudspeakers, the velocity vector points towards the same direction as intended by VBAP.

Energy Vector
Following the idea of the velocity vector, the energy vector $\boldsymbol{r}_{\mathrm{E}}$ [35, 36] was defined as

$$\boldsymbol{r}_{\mathrm{E}} = \frac{\sum_{l=1}^{\mathrm{L}} g_l^2 \boldsymbol{\theta_l}}{\sum_{l=1}^{\mathrm{L}} g_l^2}. \tag{5}$$

This model assumes an energetic superposition of the loudspeaker signals and is expected to model the localization direction for higher frequencies or broadband signals. The magnitude of the energy vector can also used to describe spatial distribution of energy [17] and the perceived width of phantom sources [37].

4.2. Prediction of the Experimental Results

Along with the experimental results, Figures 2 to 5 show the different predictions. Obviously, the direction of the velocity vector is identical to the desired panning direction for all evaluated panning methods due to the regular loudspeaker arrangement. For both Ambisonics variants, it is also identical to the direction of the energy vector. The direction of the intensity vector is very close to the one of the velocity vector. This finding shows that the intensity vector is the measured counterpart of the velocity vector, even under non-free-field conditions.

Table 3: *Average absolute deviation of predictions from median experimental results for different panning methods.*

	VBAP	MDAP	basic	max-r_{E}
Binaural Model	2.35°	3.07°	4.92°	3.37°
Intensity Vector	2.75°	2.12°	1.89°	2.41°
Velocity Vector	2.35°	1.28°	1.58°	1.05°
Energy Vector	1.60°	1.44°	1.58°	1.05°

Table 3 compares the deviation of the predictions from the median experimental results. The binaural model yields the worst prediction, especially for basic Ambisonics. Better results are achieved by the intensity vector. The vector models yield the smallest deviations from the experimental results. In detail, the energy vector predicts the VBAP localization better, as it includes the effect that the localization tends towards the loudspeakers.

The deviations of the predictions from the experimental results are similar to the standard deviations of the experimental results. Thus, all predictors seem to be suitable for the localization of frontal phantom sources at the central listening position. However, it is remarkable that the simplest models yield the best predictions at the same time and the most complex model the worst predictions.

5. CONCLUSION

This contribution investigated frontal phantom source localization at the central listening position using VBAP, MDAP, basic, and max-r_{E} Ambisonics on a circle of 8 loudspeakers. The match between the median experimental results and the desired panning direction was best for max-r_{E} Ambisonics and worst for VBAP. However, the standard deviation of VBAP was the smallest of all panning methods. Obviously, the standard deviation increases with the number of active loudspeakers. This is expected to be even more relevant for off-center listening positions [2, 5, 38]. The standard deviation was found to be dominated by the inter-subjective standard deviation, i.e. the differences between the subjects.

The experimental results were compared to a binaural localization model, the measured intensity vector, and the velocity and energy vectors. All these predictors seemed to be suitable for the prediction of the experimental results. It is remarkable that the velocity and energy vectors as the simplest predictors yielded the best predictions at the same time. This finding justifies the use of these predictors in practice. Moreover, there exist first hints that they can be also applied to vertical or three-dimensional amplitude panning [39]. However, their applicability for lateral phantom sources or off-center listening positions is still under investigation.

6. ACKNOWLEDGMENTS

The author thanks all subjects for their participation and the reviewers for their helpful comments. This work was partly supported by the projects AAP and ASD, which are funded by Austrian ministries BMVIT, BMWFJ, the Styrian Business Promotion Agency (SFG), and the departments 3 and 14 of the Styrian Government. The Austrian Research Promotion Agency (FFG) conducted the funding under the Competence Centers for Excellent Technologies (COMET, K-Project), a program of the above-mentioned institutions.

7. REFERENCES

[1] Klaus Wendt, *Das Richtungshören bei der Überlagerung zweier Schallfelder bei Intensitäts- und Laufzeitstereophonie*, Ph.D. thesis, RWTH Aachen, Germany, 1963.

[2] Matthias Frank, *Phantom Sources using Multiple Loudspeakers in the Horizontal Plane*, Ph.D. thesis, University of Music and Performing Arts Graz, Austria, 2013.

[3] Eric Benjamin, Aaron Heller, and Richard Lee, "Localization in Horizontal-Only Ambisonic Systems," in *Audio Engineering Society Convention 121*, 10 2006.

[4] Stéphanie Bertet, Jérôme Daniel, Laëtitia Gros, Etienne Parizet, and Olivier Warusfel, "Investigation of the Perceived Spatial Resolution of Higher Order Ambisonics Sound Fields: A Subjective Evaluation Involving Virtual and Real 3D Microphones," in *Audio Engineering Society Conference: 30th International Conference: Intelligent Audio Environments*, 3 2007.

[5] Matthias Frank and Franz Zotter, "Localization experiments using different 2D Ambisonics decoders," in *25. Tonmeistertagung, Leipzig*, 2008.

[6] Gavin Kearney, Enda Bates, Frank Boland, and Dermot Furlong, "A comparative study of the performance of spatialization techniques for a distributed audience in a concert hall environment," in *Audio Engineering Society Conference: 31st International Conference: New Directions in High Resolution Audio*, 6 2007.

[7] Jens Blauert, *Spatial Hearing*, MIT Press, 1983.

[8] David Moore and Jonathan Wakefield, "A design tool to produce optimized ambisonic decoders," in *Audio Engineering Society Conference: 40th International Conference: Spatial Audio: Sense the Sound of Space*, 10 2010.

[9] Ville Pulkki, "Virtual sound source positioning using vector base amplitude panning," *J. Audio Eng. Soc*, vol. 45, no. 6, pp. 456–466, 1997.

[10] D. M. Leakey, "Some measurements on the effects of interchannel intensity and time differences in two channel sound systems," *The Journal of the Acoustical Society of America*, vol. 31, no. 7, pp. 977–986, 1959.

[11] Alan D. Blumlein, "British patent specification 394,325 (improvements in and relating to sound-transmission, sound-recording and sound-reproducing systems)," *J. Audio Eng. Soc*, vol. 6, no. 2, pp. 91–98, 130, 1958.

[12] Ville Pulkki, *Spatial Sound Generation and Perception by Amplitude Panning Techniques*, Ph.D. thesis, Helsinki University of Technology, 2001.

[13] Ville Pulkki, "Uniform spreading of amplitude panned virtual sources," in *Applications of Signal Processing to Audio and Acoustics, 1999 IEEE Workshop on*, 1999, pp. 187–190.

[14] Duane H. Cooper and Takeo Shiga, "Discrete-matrix multichannel stereo," *Journal of the Audio Engineering Society*, vol. 20, no. 5, pp. 346–360, 1972.

[15] Michael A. Gerzon, "With-height sound reproduction," *Journal of the Audio Engineering Society*, vol. 21, pp. 2–10, 1973.

[16] David G. Malham and Anthony Myatt, "3D Sound Spatialization using Ambisonic Techniques," *Computer Music Journal*, vol. 19, no. 4, pp. 58–70, 1995.

[17] Jérôme Daniel, *Représentation de champs acoustiques, application à la transmission et à la reproduction de scénes sonores complexes dans un contexte multimédia*, Ph.D. thesis, Université Paris 6, 2001.

[18] Mark A. Poletti, "A Unified Theory of Horizontal Holographic Sound Systems," *J. Audio Eng. Soc*, vol. 48, no. 12, pp. 1155–1182, 2000.

[19] Frank Zotter, Hannes Pomberger, and Markus Noisternig, "Energy-Preserving Ambisonic Decoding," *Acta Acustica united with Acustica*, vol. 98, no. 1, pp. 37–47, 2012.

[20] Franz Zotter and Matthias Frank, "All-round ambisonic panning and decoding," *J. Audio Eng. Soc*, vol. 60, no. 10, pp. 807–820, 2012.

[21] ITU, "ITU-R BS.1116-1: Methods for the subjective assessment of small impairments in audio systems including multichannel sound systems," 1997.

[22] Matthias Frank, Ludwig Mohr, Alois Sontacchi, and Franz Zotter, "Flexible and Intuitive Pointing Method for 3-D Auditory Localization Experiments," in *Audio Engineering Society Conference: 38th International Conference: Sound Quality Evaluation*, 6 2010.

[23] Alois Sontacchi, Hannes Pomberger, and Robert Höldrich, "Recruiting and evaluation process of an expert listening panel," in *Fortschritte der Akustik, NAG/DAGA*, Rotterdam, 2009.

[24] Matthias Frank and Alois Sontacchi, "Performance review of an expert listening panel," in *Fortschritte der Akustik, DAGA*, Darmstadt, 2012.

[25] Ville Pulkki, "Localization of amplitude-panned virtual sources II: Two- and three-dimensional panning," *J. Audio Eng. Soc*, vol. 49, no. 9, pp. 753–767, 2001.

[26] Florian Wendt, Matthias Frank, and Franz Zotter, "Application of localization models for vertical phantom sources," in *Fortschritte der Akustik, AIA-DAGA*, Meran, 2013.

[27] Laurent S. R. Simon, Russell Mason, and Francis Rumsey, "Localization curves for a regularly-spaced octagon loudspeaker array," in *Audio Engineering Society Convention 127*, 10 2009.

[28] Werner Lindemann, "Extension of a binaural cross-correlation model by contralateral inhibition. I. Simulation of lateralization for stationary signals," *The Journal of the Acoustical Society of America*, vol. 80, no. 6, pp. 1608–1622, 1986.

[29] Werner Lindemann, "Extension of a binaural cross-correlation model by contralateral inhibition. II. The law of the first wave front," *The Journal of the Acoustical Society of America*, vol. 80, no. 6, pp. 1623–1630, 1986.

[30] Brian C. J. Moore, Robert W. Peters, and Brian R. Glasberg, "Auditory filter shapes at low center frequencies," *The Journal of the Acoustical Society of America*, vol. 88, no. 1, pp. 132–140, 1990.

[31] Lloyd A. Jeffress, "A place theory of sound localization," *Journal of comparative and physiological psychology*, vol. 41, no. 1, pp. 35–39, 1948.

[32] Frederic L. Wightman and Doris J. Kistler, "The dominant role of low-frequency interaural time differences in sound localization," *The Journal of the Acoustical Society of America*, vol. 91, no. 3, pp. 1648–1661, 1992.

[33] Juha Merimaa, "Energetic sound field analysis of stereo and multichannel loudspeaker reproduction," in *Audio Engineering Society Convention 123*, 10 2007.

[34] Richard C. Heyser, "Instantaneous intensity," in *Audio Engineering Society Convention 81*, 11 1986.

[35] Y. Makita, "On the directional localization of sound in the stereophonic sound field," Tech. Rep., EBU, Review Part A, 73, 1962.

[36] Michael A. Gerzon, "General metatheory of auditory localisation," in *Audio Engineering Society Convention 92*, 3 1992.

[37] Matthias Frank, "Source width of frontal phantom sources: Perception, measurement, and modeling," *Archives of Acoustics*, vol. 38, no. 3, pp. 311–319, September 2013.

[38] Peter Stitt, Stéphanie Bertet, and Maarten van Walstijn, "Perceptual investigation of image placement with ambisonics for non-centred listeners," in *DAFx-13*, Maynooth, Ireland, September 2013.

[39] Florian Wendt, Matthias Frank, and Franz Zotter, "Amplitude panning with height on 2, 3, and 4 loudspeakers," in *International Conference on Spatial Audio*, Erlangen, Germany, 2014.

INTERACTIVE ACOUSTIC VIRTUAL ENVIRONMENTS USING DISTRIBUTED ROOM ACOUSTIC SIMULATIONS

Frank Wefers, Jonas Stienen, Sönke Pelzer, Michael Vorländer

Institute of Technical Acoustics,
RWTH Aachen University,
Aachen, Germany
{fwe,jst,spe,mvo}@akustik.rwth-aachen.de

ABSTRACT

This publication presents how the computational resources of PC clusters can be used to realize low-latency room acoustic simulations for comprehensive virtual scenes. A benefit is not only to realize more complex scenes (including a multitude of sound sources, acoustically coupled-rooms with sound transmission), but also to minimize the system response times for prototyping applications (e.g. interactive change of materials or geometry) in simpler applications.

PC clusters prove to be a suitable platform for room acoustic simulations, as the incorporated algorithms, the image source method and stochastic ray-tracing, are largely free of data interdependencies. For the computation in massive parallel systems the simulation of a room impulse response is separated into individual parts for the direct sound (DS), early reflections (ER) and diffuse late reverberation (LR). Additional decomposition concepts (e.g. individual image sources, frequency bands, sub volumes) are discussed. During user interaction (e.g. movement of the sources/listeners) the system is continuously issued new simulation tasks. A real-time scheduler decides on significant updates and assigns simulation tasks to available cluster nodes. Thereby the three simulation types are processed with different priorities. The multitude of (asynchronously) finished simulation tasks is transformed into room impulse responses. Convolution with the audio signals is realized by non-uniformly partitioned convolution in the frequency domain. The filter partitioning is adapted to the update rates of the individual impulse response parts (DS, ER,LR). Parallelization strategies, network protocols and performance figures are presented.

1. INTRODUCTION

Nowadays fast algorithms together with powerful computers allow to simulate the acoustics in rooms in almost *real-time* compatible rates [1, 2, 3, 4]. This makes comprehensive simulation of the acoustics for interactive virtual environments possible. Such systems are valuable tools for architectural prototyping, room acoustic engineering and noise assessment. The computational demands for the required simulations are very high and their comprehensiveness is usually limited by the engaged hardware. Established geometrical acoustics (GA) methods, such as image source or ray tracing methods, are efficient enough to bring down computation times within the range of a few hundred milliseconds. Advanced aspects however, in particular diffraction modeling, are exceedingly complex and make the computation several magnitudes slower. Wave-based simulation techniques can nowadays be realized in real-time for a limited low-frequency range (e.g. finite-difference time-domain (FDTD) and derivatives using graphic processors [5, 6]). Due to their time and memory complexity they do not provide a suitable solution for the full audible frequency range (20-20.000 Hz).

Figure 1: *Interactive room acoustics demo of a medieval church in the aixCAVE virtual environment at RWTH Aachen University.*

Parallelism became a key concept in achieving real-time capability of the simulations. Wave-based solvers inhere distinct data inter-dependencies. Domain decomposition approaches for numerical methods require the exchange of sound field variables (e.g. pressure) on the boundaries of spatial sub domains. These *'updates'* have to be performed for every simulation step, creating excessive amounts of communication. A shared memory access with high throughput and low latency (e.g. video RAM) is preferable here. Typical GA computations, such as image sources and ray tracing (RT) are largely independent, thus permitting isolated computation of fractions of a simulation result. Audiblity checks of image source (IS) and the tracing of rays can be easily distributed, i.e. on several cores of a single computer unit. Only the assembly of the resulting filter (e.g. based on a list of audible image sources or energy histograms) depends on all fractions of the simulation. The independence of data and partial steps in the computation make GA algorithms particularly suitable for distributed simulation on computer clusters. Given that a single shared-memory machine cannot provide the necessary performance, the use of multiple units is a promising approach to increase the required computation power. Strategies to use PC clusters for efficient *real-time* room acoustics simulations are examined in this paper.

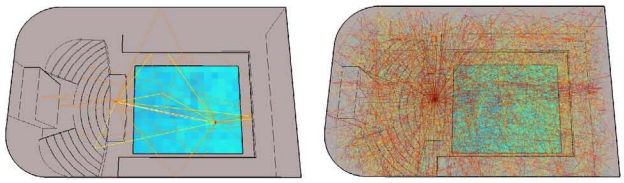


Figure 2: *Visualization of computed sound propagation paths in a concert hall model, showing image source traces (left) and simplified ray tracing (right, with very low number of rays).*

2. ROOM ACOUSTICS SIMULATION

Today's widely used acoustics simulation programs mostly base on the principles of classical GA. Already in the 1960s, Schroeder and Krokstad applied the ray tracing technique to predict acoustics properties of specific room shapes [7, 8]. Since then, GA algorithms have continuously advanced, so that modern software can handle complex wave effects, including scattering [9] and diffraction [10, 11]. Benchmark tests certified GA simulation methods as a valuable tool for the design and prediction of room acoustics [12, 13, 14, 15]. Major uncertainties in current simulations are rather reasoned by the input data than to the pure algorithms [15].

Simulation methods

State-of-the-art implementations use hybrid combinations of both ray tracing and image sources. Vorländer [16] and Vian et al. [17] presented the basis for the cone-, beam- and pyramid-tracing dialects, e.g. [2, 18], by showing that forward tracing is a very efficient method for the modeling of early reflections in a room impulse response (RIR). During the 1990s it was shown that GA cannot solely be based on specular reflections [12] which lead to the integration of scattering with activities on the prediction, measurement and standardization of scattering and diffusion coefficients of corrugated surfaces [19]. Advances in the binaural technology enabled the incorporation of spatial attributes to room impulse responses. The key equation of the contribution of one room reflection, $\underline{H}_j$, is given in spectral domain with

$$\underline{H}_j = \frac{e^{-j\omega t_j}}{ct_j} \cdot \underline{H}_{air}(ct_j) \cdot \underline{H}_{Src}(\theta, \phi) \cdot \underline{H}_{Rec}(\vartheta, \varphi) \cdot \prod_{i=1}^{n_j} \underline{R}_i \quad (1)$$

where t_j is the reflection's delay, ωt_j the phase, $1/ct_j$ the distance law of spherical waves, $\underline{H}_{Src}$ the source directivity in source coordinates, $\underline{H}_{air}$ the low pass of air attenuation, $\underline{R}_i$ the reflection factors of the walls involved, and $\underline{H}_{Rec}$ the head-related transfer function of the sound incidence at a specified head orientation. The complete binaural RIR is composed of the direct sound and the sum of all reflections. Audible results are obtained by convolution of the binaural RIR filter with an anechoic source signal.

The computation of the filters is implemented using the simulation software Room Acoustics for Virtual Environments (RAVEN) [20]. The software features a hybrid image source and ray tracing engine that can operate under real-time conditions. This was achieved by increasing the simulation speed using concepts of spatial subdivision [21].

Spatial data structures and room models

The performance of GA methods is strongly governed by intersection tests, which consume the major part of the runtime. Spatial search concepts (binary space partitioning (BSP), bounding volume hierarchies (BVH) or spatial hashing (SH)) are commonly used techniques to accelerate the intersection testing. In case of interactive geometry modifications, the spatial data structures have to be updated before a new simulation. Performance analysis showed that the most efficient way of reacting to a geometry update is the usage of a hybrid approach including two different spatial data structures. In direct comparison, BSP trees allow faster intersection tests than SH hashing [22]. However, the update of a BSP tree needs significantly more time compared to a spatial hash table. As the early reflections involve much fewer intersection tests compared to the late diffuse reflections, they can be effectively accelerated by using SH [23] for the intersection tests during IS audibility tests, in particular if the geometry is subject to frequent modifications. For the late reflections in the RIR, our RT algorithm uses BSP [21] to reduce the number of intersection tests. This reduces the total computation time of the RT process significantly, although there is a substantial delay due to the necessary reconstruction of the BSP tree. For details of the implementation the reader is referred to [22].

In general for any of the spatial data structures there is a high dependency of the performance on the number of polygons in the scene. It is known that the 3D model for acoustics calculations can (and must) be modeled with much less details than the optical representation, as shown in Figure 3. However, it is important to define proper characteristics for the surface parameters (frequency dependent absorption and scattering coefficients). In prior investigations it was analyzed how much structural detail is needed to be included in the acoustic model. The findings indicated that for typical rooms details smaller than 0.5 m can be neglected [24].

Sound transmission and diffraction

A building acoustics module in the RAVEN framework allows the simulation and auralization of sound transmission through structures [25]. Any pair of polygons in the scene can act as an bidirectional surface source and surface receiver pair. Using appropriate transfer functions the structure born sound propagation can be accounted for, e.g. through walls, floors and doors [26].

In case of coupled volumes with small apertures, but also for large objects in rooms or generally in outdoor scenes, the sound diffraction around edges is an important wave effect and must be taken into account. The RAVEN library implements edge diffraction for the IS algorithm as well as for the RT technique. Details on the implementation are published in [10]. Especially for these very complex and computationally expensive operations, the possibilities of distributed computing play an important role, so that these effects can be rendered in real-time, too.

Simulation processing

The RAVEN software has a modular architecture. It can be used by a graphical user interface, but also as a pure network service. For the distributed real-time sound field rendering, several *simulation nodes* were configured and controlled via network. Each RAVEN node is controlled using a remote interface. A simulation is remotely executed by handing over a simulation task to a simulation node. The received simulation task is then locally translated

into an execution of a simulation algorithm. An initial simulation will consist of a) the generation of an image source cloud for the current source position, b) an audibility test to filter audible image sources for the current receiver position, and c) ray tracing for the current source and receiver positions. During the interactive simulation, only steps b) and c) are continuously executed. Each simulation runs asynchronously, blocking the node in the mean time. When the computation is finished, it returns either a set of (potential/audible) image sources or histograms containing the spectral energy envelopes of the late decay. This intermediate result is then passed to a *filter controller*, which constructs time-domain room impulse responses. The obtained impulse responses for early (IS) and late (RT) parts are superposed in time domain. Each simulation task contains a snapshot of the virtual scene. Changes to the prior state, such as translated sound sources or listeners, are efficiently implemented by translation of the image source cloud using pre-calculated translation matrices. Therefore, no reconstruction of the image source tree is necessary. For ray tracing it must be mentioned that re-calculating the spectral late decay is only necessary if the position of either the source or the receiver is subject to significant translation. In typical rooms the late part of the impulse response is diffuse and homogeneous and will only be updated if the source or receiver moved more than 1 m away from the last simulated positions. In contrast the direct sound must already be updated if a source or the listener moves a few centimeters or rotates by a few degrees. This part of the impulse response is crucial for the localization of sources and is in full attention of the listener. However, this operation is very fast as only a test for a line of sight must be performed on the scene geometry and a short filter of source directivity, receiver directivity (e.g. HRTF) and air attenuation is assembled.

Shared-memory parallelization

Operations with high computation time consumption such as the generation of the image source trees, the audibility tests of image sources and the tracing of rays are parallelized at node-level using *OpenMP* [27]. For IS the parallelization is very effective in the construction of the IS cloud. Here, higher IS orders dependent on lower order image sources, but the IS tree branches of all first order sources can be constructed in parallel without any interdependencies. Especially the ray tracing algorithm scales nearly perfectly linear to the number of locally available cores and has no data interdependencies. In typical rooms, 10k-100k rays are to be shot, each with similar computation time and completely independent of each other. Therefore, all available processor cores are set up

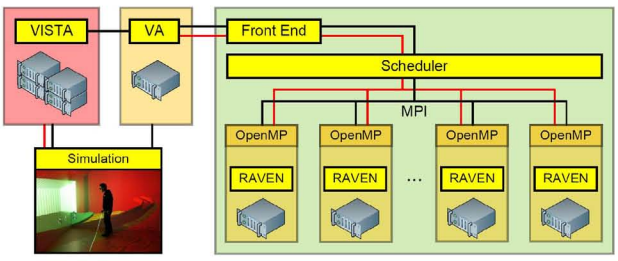


Figure 4: *Hybrid inter-node and intra-node level parallelism inside the RAVEN simulation framework.*

Simulation	Computation time (single node)
Direct sound (DS)	3.4 ms
Early reflections (ER)	23 ms
Late reverberation (LR)	2.04 s

Table 1: Average computation times for different simulation types computed on a single cluster node (shared-memory parallelization). Each runtime corresponds to the generation of a binaural filter (DS, ER or LR) for one source↔listener pair.

with one single thread on each core that traces one single ray at a time. By sorting the angles of the launched rays to spatially coherent directions the utilization of the CPU's low-level data caching is improved.

In RT different frequency bands are usually simulated separately (mostly in octave resolution) and are executed in sequential order. A parallelization across frequency bands has the disadvantage that higher frequency bands have much shorter computation times than low frequencies due to typical material and air absorption properties. In this case the workload is not equally balanced across the cores/threads leading to unnecessary overhead. Given that only few simulations are computed in parallel, the capabilities of a compute cluster, featuring a multitude of nodes cannot be fully utilized. In order to achieve the best performance, the multitude of rays must be distributed across the nodes. The energy histograms of each partial result can simply be superposed and are identical to a simulation with the summed number of rays on a single machine. Locally on each node, the room acoustics simulation should employ shared memory parallelization due to their excellent scaling capabilities, instead of running multiple concurrent simulators on a multi-processor node. The hybrid inter-node and intra-node level parallelization concept considered in this paper is illustrated in figure 4 .

Test scene

As a representative test case, a virtual scene of the medieval spanish church *San Juan de Baños* has been selected to benchmark the performance of distributed simulations of room acoustics. The virtual and acoustic model are depicted in Figure 3. Image sources are calculated up to the second order. Ray tracing has been performed using a target filter length of 2.8 seconds with 10.000 particles for each of the 10 frequency bands in octave resolution from 31 Hz to 16 kHz, a time resolution of 6 ms and a detection sphere radius of 0.5 m. For accelerating simulation time, the BSP method has been used. Furthermore, the computations were parallelized using 24 threads (compare section 5 for further details). The resulting computation times on a *single computer* node of the cluster are listed in table 1. The technical details of the hardware are outlined in section 5.

3. AUDIO RENDERING

Audio rendering is considered the process of transforming the monaural source signals into appropriate listener signals, which in our case are binaural. The foundation is the description of the virtual scene, referred to as the *scene state*. It describes all stationary attributes of acoustically relevant entities in the virtual scene. These include positions, orientations, velocities of objects and specific properties, like directivity of sources, head-related impulse

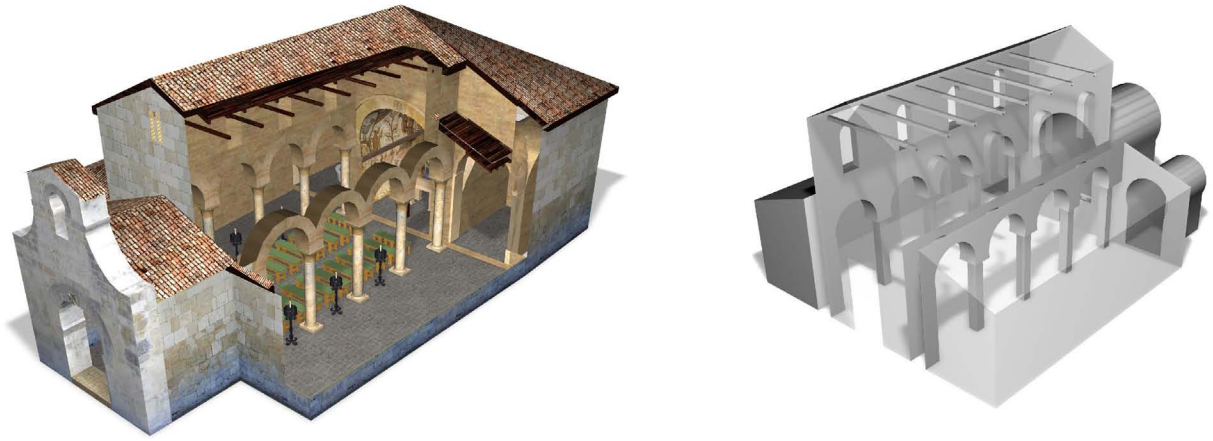


Figure 3: *Visual and acoustics model of San Juan de Baños chapel. The acoustic model (right) uses* 1.300 *polygons with an estimated volume of* $756\,m^3$ *and surface of* $776\,m^2$*. The predicted reverberation time (Sabine) is* 3 *seconds.*

response (HRIR) datasets of listeners. Moreover, the properties of the propagation medium (static pressure, temperature, humidity), the scene geometry (polygons), acoustic parameters of surfaces (absorption, scattering), etc. Any modification, for instance pose changes acquired from motion tracking that control the movement of the listener, result in a state change. User interaction may additionally alter positions, orientations or even change internal states of objects (switching material of a wall, interacting with a machine that results in a change of emitted sound). Also, autonomous objects (e.g. avatars or moving vehicles) can be controlled independently by a Virtual Reality (VR) software. Given a scene change (update), the auralization software needs to rapidly update the acoustic stimuli for the user, with respect to the new situation.

Most auralization methods describe the virtual scene only in an instant of time disregarding history data—all computation is based on the current state from the viewpoint of the listener. This simplification is acceptable as long as sound propagation times are negligible small, i.e. in small spaces. For the simulation of room acoustics and outdoor scenarios it is helpful to preserve access to past states of the scene along a time history. This allows to provide for acoustic effects that occur with respect to finite medium propagation (i.e. delayed arrival of acoustic events, Doppler shifts at sound source) and, in addition, enables algorithms that can reuse results, i.e. a ray tracing simulation for a specific position. This becomes even more relevant, when several hundred simulation jobs are executed in parallel, yet the delivery of results is not fast enough to cover a specified amount of spatial resolution. In this case, the contribution of an obsolesce simulation result may lead to a better auralization of the current scene state. In the following we present a scene data structure that fulfills this requirements and suits a real-time computation.

Scene description

A single *scene state* of a virtual environment describes a static snapshot of all relevant entities and their properties. It represents the top-level (global scene) and consists of multiple sub-states (e.g. states of sources and listeners). These on their own are assembled by further lower level states (e.g. position or orientation). States form a hierarchy and are implemented as a tree. Each time the virtual scene is altered, a new scene state is derived from the current state (copy). Then only the affected parts (sub trees or leafs) actually modified in the newly derived state. It holds the differential information to the prior state. Unchanged elements are simply referenced, keeping the memory footprint compact. A state is kept in memory, until the last reference is removed. This scheme makes it very easy to compare different states and to track back parameters over time. Moreover, the data structure allows to handle asynchronous resource allocations and deallocations. In detail, the creation of states is taking place in the control-flow of the caller (external user), while destruction of states is decided in the context of the real-time audio rendering (in case a state becomes obsolete). Decoupling of these actions is realized using object pools and reference counters. Thereby lock-free data structures are employed to avoid the problem of blocking in the time-critical audio processing.

Update rates in VR applications are often determined by connected input devices (e.g. motion tracking with 120 Hz). For the signal processing some parameters, like positions of objects, are required with even higher rates, i.e. if the audio streaming process updates its elements with 340 frames per second (128 samples at 44.100 Hz). Parameter interpolation [28] can be easily added on top of the scene description. In particular, positions of fast moving objects are real-time predicted and interpolated using a motion model, producing a quasi-continuous trajectory. Attributes can be queried, even at audio sampling rates, if necessary.

Signal processing

Signal processing is based on the abstraction of propagation of sound between a single source and a single listener. The superposition of all source↔listener pairs represent the entire virtual situation, and their relationship is illustrated in figure 5. The core element within the processing chain is represented by the convolution module that convolves the source signal with the simulated RIR. For the sake of high-quality auralization that minimizes acoustic artifacts, some aspects like propagation delay and high resolution distance-dependent attenuation caused by spreading loss have to be realized independently. In the following the processing is explained in more detail.

Common simulation methods encode the traveling time $T = r/c$ of the sound waves into the impulse response, as it is the case

for a measured impulse response. The direct sound impulse is located at the offset $T \cdot f_S$ (samples). Prior filter coefficients are zero, realizing a delay that corresponds to the traveling time. When it comes to real-time auralization, this concept has several disadvantages: firstly, leading zeros increase the computational effort of the filtering procedure, and secondly, changes in distance can only be updated with each filter exchange. Also they manifest in time-shifts of parts of the impulse response, which is inconvenient for the fast convolution. Simple implementations based on simple switching of filter coefficients or applying additional cross fading of the output stream may have negative effects: they usually produce audible artifacts (clicks, comb-filters). A superior approach is to separate propagation delays from the RIRs by the use of a variable delay-line (VDL). Based on the continuous-time motion model mentioned earlier the VDL parameters can be continuously adapted (per sample) and hence comb filter artifacts can be avoided. VDL can be realized with a low computational complexity (i.e. with fractional delays or polynomial interpolators).

For the use with delay-lines, the initial traveling time is removed from the simulated room impulse responses (see figure 6). The direct sound impulse (red) is always located at the beginning (accounting for a fixed pre-ring offset in the case of non-minimum-phase HRIRs). The other parts, ER (green) and LR (blue), reside in more or less fixed ranges of filter coefficients. Moreover, the impulse responses are normalized in magnitude, so that the amplitude of the direct sound part is 1 (amplification of the complete impulse response by r). This approach simplifies the time-varying finite impulse response (FIR) filtering and allows the effective use of output crossfading techniques [28] to prevent audible artifacts for the filter changes.

The filtering with the room impulse response is implemented using non-uniformly partitioned convolution in the frequency-domain [29] [30] [31]. Our convolution engine performs Overlap-Save (OLS) using the Fast Fourier Transform (FFT) and is specialized for binaural processing (two output channels). The non-uniform filter partition is optimized [32] [33] with respect to the filter subdivision and desired update rates for these parts of the room impulse response (DS, ER, LR). A smooth exchange of filters is accomplished by computing the convolution result streams for both filters (old and new) during exchange and cross-fading the output samples in time-domain.

The second to last step is reapplying the previously removed gains for spreading losses to the output signals. These are now applied *per sample*, allowing to smoothly adapt the gain over the samples of each frame. Updating them *per filter* results in steps in the gain envelope, which cause clearly audible clicks due to inconstancy of the signal wave form. Again, based on the motion model, the source↔listener distances are available for the beginning and end time of each audio frame and then a gain envelope is interpolated. The process is illustrated at an example trajectory in figure 7. For frame rates >300 Hz this piece-wise linear gain curve turns out to be sufficient. Finally, the left/right ear signals of each sound source are added up (superposition at the listeners ears), resulting in a single binaural stimulus for each listener.

In sense of a more physically-correct auralization, the time-varying propagation delays should be considered individually for each image source (specular reflection), too. The relative shifts in time-delays for the direct sound and early reflections are more or less independent. Each image source up to a given order, could be implemented using an independent VDL and convolution of HRIR and reflection filter. But this would increase the computational effort of the auralization by several magnitudes, limiting the possible number of sound sources and listeners. Although it remains an open scientific question, if this expense is necessary with respect to audiblity.

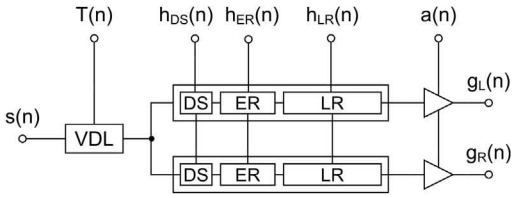


Figure 5: *Audio rendering of a source↔listener sound path. The monaural source signal $s(n)$ is transformed into a binaural room acoustic stimulus $g_{L/R}(n)$. Individual parts of the impulse response are updated asynchronously. Propagation delays and distance gains are updated per sample and realized apart from the RIR.*

Time-varying scenes

Due to the block-based signal processing, the maximum filter update rate is determined by the frame rate of the audio stream. Therefore, it is reasonable to check for filter updates and trigger them within the context (thread) of the audio processing. Embedded into this processing call, the auralization module performs parameter changes that require time-critical update rates (like distance gain and medium propagation delay). The actual filter calculation is realized in separate threads. Validation for updates and the simulation are decoupled using update requests, called *tasks*, for certain parts of the signal processing chain (i.e. the reconstruction of ER and LR). Tasks are lightweight, allowing to issue (create) them within the audio context, without diminishing the computation time. Each task includes all the necessary information for performing a room acoustic simulation (DS, ER or LR). Together with a reference tag and a unique identifier, each task is then handed over to a *scheduler*, which runs asynchronously, in a dedicated thread. It is responsible for the further handling of the request, its simulation and the final application of the computed updates. Details of the scheduler are explained in the succeeding section 4. Eventually, simulation results are received in yet another decoupled thread that assembles the entire RIR and updates relevant filter partitions in the convolution process. For the direct-sound part, this is a HRIR signal, the audibility status (audible, inaudible) and optionally a signal with diffracted components of the direct sound. The result of an image source computation is a list of audible image sources, which are combined into an image source filter. This filter is inserted into the given RIR of the according source↔listener pair and the covered range of filter coefficients is updated in the convolution. Ray tracing results (late reverberation filters) are handled in a similar way.

On many-core shared-memory systems several simultaneous simulator threads are created and assigned to specific types of tasks. The distribution of the computation (signal processing, room acoustic simulation, etc.) to the available computation units (processors or cores) is a distinct procedure that relies on the available hardware. The decision of calculating and updating sections of the transfer path depends on the number of source↔listener pairs, available cores, computational complexity, etc. Direct sound tasks are updated instantaneously within the context of an audio frame.

All other tasks are handed over to the scheduler, which decides if the task is worth an update and then assigns its computation to an available core. Predefined numbers of cores are assigned to the different computations (real-time convolution, image source and ray tracing simulations), with respect to the underlying hardware and scene complexity. Thereby the real-time convolution has the highest priority. Dropouts must be strictly avoided. The remaining cores are assigned to room acoustics simulations. Usually, a reasonable compromise between an acceptable image source computation time and ray tracing update rate has to be chosen.

4. DISTRIBUTED SIMULATION

The presented concept of distributed room acoustic simulation is realized in fashion of a client-service structure (see fig. 4). The audio renderer, as a client, generates an arbitrary amount of tasks and transfers those to scheduler. The scheduler acts as a server, planning and organizing the received simulation requests and passing on tasks to the worker nodes in the cluster. In shared-memory systems, data is exchanged between threads by the use of shared variables or pointers. Convenient tools for compiler-generated loop-level parallelization are available, for instance *OpenMP* [27]. In contrast, distributed memory systems lack a common memory which all nodes can collectively access. These systems are typically programming using the paradigm of *message passing*, e.g. using *MPI* [34]. All data has to be transmitted in-between nodes using *messages* that are delivered by communication channels. *MPI* introduces an abstraction layer, that hides underlaying hardware structures like the machine type of each node and network communication channels, making deployment, execution and expansion easily manageable. Most modern clusters are assembled from multi-/many-core computers (distributed shared-memory systems). Therefore, both parallelization concepts (e.g. MPI + OpenMP) are often combined, known as *hybrid parallelization*. This is as well applied in the presented approach.

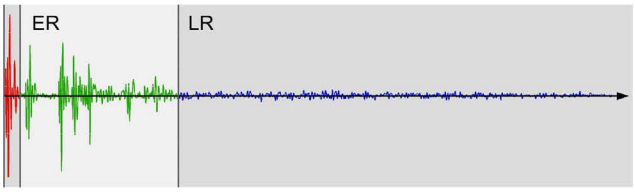


Figure 6: *Parts of simulated room impulse responses (DS, ER, LR) are mapped to individual filter segments in the non-uniformly partitioned convolution.*

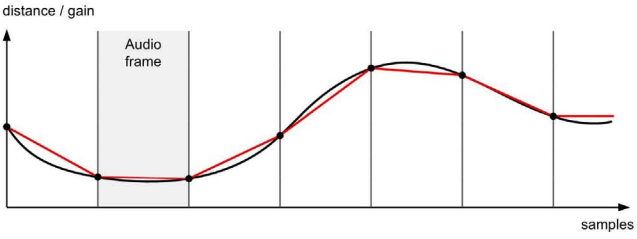


Figure 7: *Interpolation of 1/r gains for spreading losses. (Black: Continuous distance function obtained from motion model, Red: Linear-interpolated per sample, used as VDL input)*

Method

The problem of distributed room acoustic simulations particularly suit the hybrid parallelization on computer clusters because it requires minimal network communication. Given that a simulation task is scheduled for computation on a vacant worker node, only the task data (a few kilobytes) need to be transmitted. The simulation node can adapt to any given virtual scene and requested simulation type. During simulation, no further communication with any other node is necessary. Each simulation node executes simulations sequentially. Once a task has been assigned to it, the scheduler considers the worker node busy and does not send any further tasks. On completion of a task, the worker node communicates a result message to the master node containing the simulation results, i.e. the binaural filter parts. In order to decrease the package size, leading and tailing zeros of the impulse response are stripped prior to transmission. After successful transmission, the worker node receives new tasks.

Scheduling

During a running *real-time* auralization, a vast amount of simulation tasks is generated in the context of the audio rendering, exceeding the available computing power of the cluster power by far (i.e. on every scene state change for each source↔listener pair). But this design is intended . It is the responsibility of the scheduling instance to cope with the flood of simulation tasks without compromising the rendering process. The large number of updates allows the scheduler to decide which update is meaningful and which parts are given the highest priority. In the implemented approach, a two-phase scheduling procedure is applied to prevent high network load and to maintain a lightweight handover of tasks between rendering and scheduling. In order to achieve this, both the client-side and the server-side, carry out a replacement strategy with the same procedure, thus preventing transmission of tasks that most likely would have been discarded in the first place.

This paper considers a simple scheduling strategy by means of priority and replacement. An existing task is considered outdated, if an incoming task with same reference exhibits a novel time stamp, and will therefore be discarded. Tasks are treated equally, i.e. first task is served first and the list of pending tasks is polled cyclically.

5. BENCHMARKS

In order to evaluate the performance of the proposed distributed room acoustic simulation, benchmarks were conducted on a compute cluster. The throughput of simulations was measured under real-time constraints with different numbers of cluster nodes. All three different simulation types (DS, ER and LR) were benchmarked individually, with a variation of the total number of cluster nodes between 1 and 24. For the tests the scheduler was flooded with tasks from the client side, simulating real-time circumstances, where vast amounts of tasks are issued. This assured that the system was always stressed to the maximum. The scheduler then organized and planned the execution of tasks and distributed them to the available compute nodes, were they were simulated. On client side, the finished tasks collected, analyzed and an the achieved overall update rate $f_{\max}$ [filter updates per second] was counted. This measure is first of all unrelated to a number of sound source and listeners. The compute resources can be shared accordingly,

when a virtual scene contains M sources and N listeners. Given the assumption that the simulation times are mostly independent of the individual source↔listener pairs (equal sharing), the achievable update rate per sound path is approximately $f_{\max}/(M \cdot N)$.

Hardware

The benchmarks were conducted on the computing infrastructure around the *aixCAVE* CAVE-like virtual reality system (fig. 1) at RWTH Aachen University. This display system combines five-sided high-resolution back-projection screen setup with active shutter technique for stereoscopic vision. An optical tracking system captures motion of the user and of input devices thus providing the possibility to interact with the virtual scene. A dedicated acoustics sub system (auralization server) receives control information and scene data over a network interface from the VR application. This system is linked to a high performance computer cluster. Auralization server and compute cluster communicate via a custom designed TCP/IP network protocol over Gigabit-Ethernet. Cluster nodes communicate with each other using the Message Passing Interface (MPI) in form of the OpenMPI library version 1.6.4 (*multi-threaded*). Each cluster machine consists of an Intel Xeon X5650 (*Westmere*) processor with 12 cores running at 2.7 GHz and 24 GB DDR3 memory available. For the benchmark testing a division of 24 similar nodes and an additional master node of same specifications has been used exclusively. The nodes are running Scientific Linux version 6.4 as operating system with kernel version 2.6.32 (*x86_64*). The software and test suite has been implemented in the C++ programming language and was compiled and linked with g++ version 4.4.7.

Results

Figure 8 shows the achieved update rates as a function of the number of used cluster nodes for the chapel scene. The achievable maximum update rate $f_{\max}$ is affected by two aspects: (a) the runtime of the simulation on each node and (b) the data transmission times for inter-node communications. The computation times for the ray tracing on each node are in the range of 2 seconds (table 1). An almost perfect linear scaling (double the number of compute nodes ↔ double the filter update rate) can be observed. This indicates that data transmission times are comparably small. The image source computation times are magnitudes shorter. On a single node they are in the range of 20 ms. Here, the scaling is nearly linear as well, until a number of twelve nodes. But beyond that, the inter-node communication becomes a bottleneck and starts to diminish the performance benefits by additional nodes. For 20 nodes and above the limitation becomes clear. The smaller the computation time of a task is, the more significant becomes the necessary data communication for it. Direct sound audibility checks compute in less than 4 ms (table 1). For them, the scaling is far from optimal. When twelve or more nodes are employed, the communication overhead becomes a serious bottleneck and no significant improvements can be achieved by using further nodes.

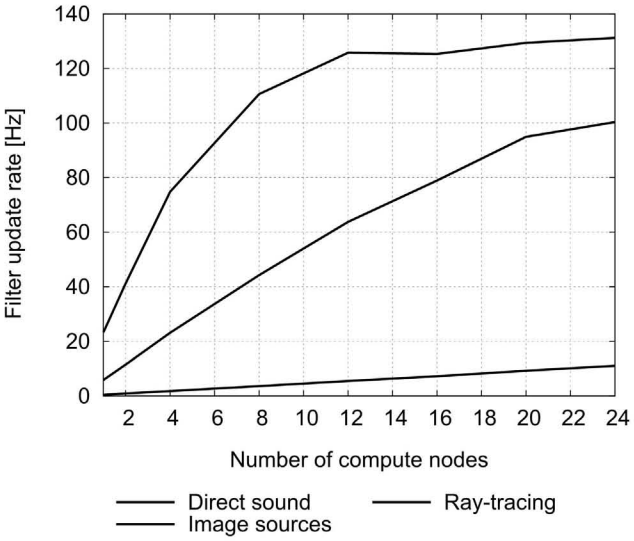


Figure 8: *Maximum filter update rates $f_{max}(N)$ for individual parts (DS, ER, LR) over the number of cluster nodes N.*

6. CONCLUSIONS

The results show a clear benefit from a distributed computation. Filter update rates can be significantly increased by several magnitudes over a simulation which runs on a single (multi-core) machine only. It can be concluded, that mostly the ER and LR tasks gain profits from the cluster implementation for distributed simulation. Short-timed computations in the range of milliseconds, like simple DS audiblity checks, are too much affected by the required data transmission. It is suggested to avoid the distribution of these tasks and compute them on the target system directly. By using more compute nodes, the update rates, at least of the ray tracing, are likely to be even higher.

7. OUTLOOK

Although the achieved filter update rates are high, the method does not minimize the latency of the computation process itself. The system could achieve > 10 ray tracing simulations per second, but still the time from task creation to finish was similar to that on a single machine (table 1). That is due to the fact that each task was only parallelized on each node and could not exploit the whole cluster. In order to speed up the single tasks and thus reduce the latencies, single IS or RT simulations must be further decomposed and distributed to multiple nodes (e.g. partitioning of image sources, different ray-packages).

For conceptually simple scenes the cluster parallelization might outreach the necessary update rates with respect to the human perception. Which rates are reasonable for the individual parts of the simulation remains to be researched. We see target applications of the proposed technique in the real-time auralization comprehensive urban noise scenarios (large numbers of sound sources and listeners) and room acoustics planning and prototyping (interactive change of materials, A-B comparisons).

8. ACKNOWLEDGMENTS

The authors thank Alexander Diaz Chyla for providing the 3D models and his work for creating the demo scene. Special thanks go to Dominik Rausch of the RWTH Virtual Reality Group for supporting this project and helping with the implementation.

9. REFERENCES

[1] L. Savioja, J. Huopaniemi, T. Lokki, and R. Väänänen, "Creating Interactive Virtual Acoustic Environments," *Journal of the Audio Engineering Society*, vol. 47, no. 9, 1999.

[2] T. A. Funkhouser, N. Tsingos, I. Carlbom, G. Elko, M. Sondhi, J. E. West, G. Pingal, P. Min, and A. Ngan, "A beam tracing method for interactive architectural acoustics.," *Journal of the Acoustical Society of America*, vol. 115, 2004.

[3] T. Lentz, D. Schröder, M. Vorländer, and I. Assenmacher, "Virtual Reality System with Integrated Sound Field Simulation and Reproduction," *EURASIP journal on advances in signal processing*, vol. 2007, 2007.

[4] D. Schröder, F. Wefers, S. Pelzer, D. S. Rausch, M. Vorländer, and T. Kuhlen, "Virtual Reality System at RWTH Aachen University," in *20th International Congress on Acoustics (ICA 2010), Sydney, Australia*, 2010.

[5] L. Savioja, "Real-time 3D finite-difference time-domain simulation of low-and mid-frequency room acoustics," in *Conference on Digital Audio Effects (DaFX-10), Graz, Austria*, 2010.

[6] N. Raghuvanshi, R. Narain, and M.C. Lin, "Efficient and Accurate Sound Propagation Using Adaptive Rectangular Decomposition," *IEEE Transactions on Visualization and Computer Graphics*, vol. 15, no. 5, 2009.

[7] M. Schroeder and B. Atal, "Computer simulation of sound transmission in rooms," *Proceedings of the IEEE*, vol. 51, pp. 536 – 537, 1963.

[8] S. S. A. Krokstad and S. Sorsdal, "Calculating the acoustical room response by the use of a ray tracing technique," *Journal of Sound and Vibration*, vol. 8, pp. 118–125, 1968.

[9] D. Schröder and A. Pohl, "Modeling (Non-)uniform scattering distributions in geometrical acoustics," *Proceedings of Meetings on Acoustics*, vol. 19, no. 1, 2013.

[10] D. Schröder and A. Pohl, "Real-time hybrid simulation method including edge diffraction," *EAA Auralization Symposium, Espoo, Finland*, 2009.

[11] D. Schröder, P. Svensson, U. Stephenson, A. Pohl, and M. Vorländer, "On the accuracy of edge diffraction simulation methods in geometrical acoustics," in *Proc. of InterNoise 2012, New York*, 2012.

[12] M. Vorländer, "International Round Robin on room acoustical computer simulations," in *15th International Congress on Acoustics*, Trondheim, Norway, 1995.

[13] I. Bork, "A comparison of room simulation software - the 2nd round robin on room acoustical computer simulation," *Acustica - Acta Acustica*, vol. 86, 2000.

[14] I. Bork, "Report on the 3rd round robin on room acoustical computer simulation - part ii: Calculations," *Acta Acustica united with Acustica*, vol. 91, 2005.

[15] S. Pelzer, M. Aretz, and M. Vorländer, "Quality assessment of room acoustic simulation tools by comparing binaural measurements and simulations in an optimized test scenario," *Acta acustica united with Acustica*, vol. 97, no. S1, 2011.

[16] M. Vorländer, "Simulation of the transient and steady state sound propagation in rooms using a new combined sound particle - image source algorithm," *Journal of the Acoustical Society of America*, vol. 86, 1989.

[17] J. Vian and D. van Maercke, "Calculation of the room impulse response using a ray-tracing method," *Proceedings of the ICA Symposium on Acoustics and Theatre Planning for the Performing Arts, Vancouver, Canada*, 1986.

[18] U. Stephenson, "Quantized pryamidal beam tracing - a new algorithm for room acoustics and noise immission prognosis," *ACTA ACUSTICA united with ACUSTICA*, vol. 82, 1996.

[19] P. D. Antonio J. J. Embrechts J. Y. Jeon E. Mommertz T. J. Cox, B.-I. L. Dalenbäck and M. Vorländer, "A tutorial on scattering and diffusion coefficients for room acoustic surfaces," *Acta Acustica united with ACUSTICA*, vol. 92, 2006.

[20] D. Schröder, *Physically Based Real-time Auralization of Interactive Virtual Environments*, Ph.D. thesis, RWTH Aachen University, 2011.

[21] D. Schröder and T. Lentz, "Real-time processing of image sources using binary space partitioning," *Journal of the Audio Engineering Society*, vol. 54, no. 7/8, 2006.

[22] S. Pelzer, L. Aspöck, M. Vorländer, and D. Schröder, "Interactive real-time simulation and auralization for modifiable rooms," *Proceeding of the International Symposium on Room Acoustics, Toronto, Canada*, 2013.

[23] A. Ryba D. Schröder and M. Vorländer, "Spatial data structures for dynamic acoustic virtual reality," *Proceedings of the 20th International Congress on Acoustics (ICA), Sydney, Australia*, 2010.

[24] S. Pelzer and M. Vorländer, "Frequency- and time-dependent geometry for real-time auralizations," *Proceedings of 20th International Congress on Acoustics*, 2010.

[25] F. Wefers and D. Schröder, "Real-time auralization of coupled rooms," Espoo, Finland, 2009.

[26] R. Thaden, *Auralisation in Building Acoustics*, Ph.D. thesis, RWTH Aachen University, Aachen, Germany, 2005.

[27] OpenMP, "Application program interface specifications version 4.0," API, July 2013.

[28] J.-M. Jot, V. Larcher, and O. Warusfel, "Digital Signal Processing Issues in the Context of Binaural and Transaural Stereophony," in *Audio Engineering Society Convention 98*, Feb 1995.

[29] W. G. Gardner, "Efficient convolution without input-output delay," *Journal of the Audio Engineering Society*, vol. 43, 1995.

[30] G. P. M. Egelmeers and P. C. W. Sommen, "A New Method for Efficient Convolution in Frequency Domain by Nonuniform Partitioning for Adaptive Filtering," *IEEE Transactions on Signal Processing*, vol. vol 44, 1996.

[31] E. Battenberg and R. Avizienis, "Implementing Real-Time Partitioned Convolution Algorithms on Conventional Operating Systems," in *Digital Audio Effects Conference (DAFx)*, 2011.

[32] G. Guillermo, "Optimal filter partition for efficient convolution with short input/output delay," in *Audio Engineering Society, Convention Paper 5660*, 2002.

[33] F. Wefers and M. Vorländer, "Optimal filter partitions for non-uniformly partitioned convolution," in *45th AES Conference on Applications of Time-Frequency Processing in Audio, Helsinki, Finland*, 2012.

[34] MPI, "A message passing interface standard version 2," API.

OPTIMIZED SPHERICAL SOUND SOURCE FOR AURALIZATION WITH ARBITRARY SOURCE DIRECTIVITY

Johannes Klein
Institute of Technical Acoustics,
RWTH Aachen University
Aachen, Germany
jck@akustik.rwth-aachen.de

Martin Pollow
Institute of Technical Acoustics,
RWTH Aachen University
Aachen, Germany
mpo@akustik.rwth-aachen.de

Michael Vorländer
Institute of Technical Acoustics
RWTH Aachen University
Aachen, Germany
mvo@akustik.rwth-aachen.de

ABSTRACT

The auralization of measured room impulse responses (RIRs) is traditionally bound to the directivity of the source as well as of the receiver. For the comparability of room acoustical measurements ISO 3382 requires the source and the receiver to be of an omni-directional directivity. Other source directivity patterns cannot be auralized using RIRs obtained this way.

In order to include the spatial information the room impulse response has either to be measured with a sound source of the desired directivity or - assuming the room to be a linear time-invariant system - it can be generated by superposing a set of measurements with a source of known directivity. The advantage of the latter method is that it generates a set of RIRs that can be used to derive the RIR for an arbitrary directivity up to a certain spherical harmonic order in post processing.

This article describes a superposition method and a specialized measurement source for the measurement of room impulse responses for an arbitrary source directivity and discusses their capabilities and the limitations. The measurement source was developed using an analytical model. The directivity patterns used for the post processing originate from high-resolution measurements of the actual device. The deviation compared to the analytical model is analyzed regarding the radiation pattern and the achievable synthesis accuracy.

1. INTRODUCTION

The results of room impulse response measurements are inextricably linked with the directivity of the employed sources and receivers. To ensure the comparability of measured standardized room acoustical parameters, ISO 3382 requires the sources and receivers to have an omni-directional directivity [1]. By excluding the influence of any directivity it neglects important information for realistic auralizations and room acoustical analysis besides the standardized parameters [2].

A sequential synthesis method employing an optimized measurement source was developed in previous research [2, 3, 4]. The method allows for the synthesis of room transfer functions of sources with an arbitrary directivity in hindsight of an extensive measurement. It complies with the requirements of the ISO 3382 and simultaneously gathers all source directivity related information about the room.

Conventional measurement sources such as dodecahedron loudspeakers are not well suited for the required measurements [4]. The optimized measurement source was developed to provide the required radiation features and to speed-up the measurement process. During the development of the optimized source its directivity was simulated using an analytical model of a vibrating cap on a sphere [5].

For room acoustical applications the synthesis method and source are described in [6]. This article focuses on the comparability of the real source with the analytical model used during the development.

2. SYNTHESIS OF ROOM TRANSFER FUNCTIONS

The *target room transfer function* for a source with a certain *target directivity* can be synthesized by superposing single room transfer functions obtained in several physical orientations with a measurement source of known directivity [2, 3]. Using a large set of orientations greatly enhances the spatial resolution of the possible target directivity patterns. The weights for the superposition can be derived from the known target directivity and the measurement source directivity in the spherical harmonics domain. The required computational steps have to be executed separately for every frequency. To enhance the readability of the equations the frequency dependence is omitted in this article.

2.1. Spherical Harmonics

In all further considerations, ϑ and φ are the elevation and the azimuth angle of a spherical coordinate system with r being its radius. Two-dimensional square-integrable functions $f(\vartheta, \varphi)$ on the surface of a unit sphere in $\Re^3$ can be represented using *spherical*

harmonics. The complex functions

$$Y_n^m(\vartheta,\varphi) = \sqrt{\frac{(2n+1)}{4\pi}\frac{(n-m)!}{(n+m)!}} \cdot P_n^m\left(\cos\vartheta\right) \cdot e^{\mathrm{j}m\varphi} \quad (1)$$

span the space of scalar functions on the unit sphere [5]. Herein, P_n^m is the *associated Lengendre function* of the first kind of the m^{th} degree in the n^{th} order [7].

The functions Y_n^m can be weighted with individual coefficients $\hat{f}_n^m$ and superposed to yield the shape of the directional function $f(\vartheta,\varphi)$ in an operation called *spherical harmonic expansion* [5]

$$f(\vartheta,\varphi) = \sum_{n=0}^{\infty}\sum_{m=-n}^{n} \hat{f}_n^m \cdot Y_n^m(\vartheta,\varphi). \quad (2)$$

To obtain the coefficients for a function the *spherical harmonic transform*

$$\hat{f}_n^m = \oint_{S^2} f(\vartheta,\varphi) \cdot \overline{Y_n^m(\vartheta,\varphi)}\,\mathrm{d}\Omega \quad (3)$$

has to be performed [5].

The coefficients can be stored consecutively in a *coefficient vector* $\hat{\mathbf{f}}$, the spherical harmonics in the *function matrix* $\mathbf{Y}$ and the sampled values of the spatial function in a *value vector* $\mathbf{f}$. The expansion and transform are simplified to matrix multiplications

$$\mathbf{f} = \mathbf{Y} \cdot \hat{\mathbf{f}} \quad (4)$$

$$\hat{\mathbf{f}} = \mathbf{Y}^{+} \cdot \mathbf{f}, \quad (5)$$

with $\mathbf{Y}^{+}$ being a generalized inverse of $\mathbf{Y}$ resulting in a not generally unique and exact result for the transform.

A spherical harmonic transformed function can be rotated by an angle α about the z-axis by multiplication with the *Euler rotation term* $e^{-\mathrm{j}m\alpha}$. Rotations about the y-axis require a full *Wigner-D rotation* [8].

Spherical harmonics offer a unified description of a sound source directivity regardless of the distribution of the measurement points. Combined with its efficient rotation, this makes spherical harmonics the calculation method of choice for the synthesis method presented here.

2.2. Synthesis Method

The single room transfer functions obtained with the measurement source in all physical orientations O in a specific acoustical environment are stored in the *frequency response vector*

$$\mathbf{h} = \left[h_1, h_2, \ldots, h_O\right]. \quad (6)$$

The goal is to synthesize the room transfer function for the target directivity. Therefore, the single room transfer functions are superposed applying a *weighting vector* $\mathbf{g}_T$ resulting in the room transfer function

$$h_T = \mathbf{h} \cdot \mathbf{g}_T, \quad (7)$$

for the desired target directivity. The directivity of the measurement source can be described in a *directivity matrix*

$$\hat{\mathbf{D}} = \left[\hat{\mathbf{d}}_1\,\hat{\mathbf{d}}_2 \cdots \hat{\mathbf{d}}_O\right], \quad (8)$$

containing the respective SH-coefficients of the directivity of the measurement source in all physical orientations O as column vectors. A synthesized directivity coefficient vector

$$\hat{\mathbf{d}} = \hat{\mathbf{D}} \cdot \mathbf{g} \quad (9)$$

can be generated by weighting and superposing the single directivity columns of the directivity matrix. For a given *target directivity* $\hat{\mathbf{d}}_T$ the weighting vector

$$\mathbf{g}_T = \hat{\mathbf{D}}^{+} \cdot \hat{\mathbf{d}}_T \quad (10)$$

can be found by multiplication with the generalized inverse of the directivity matrix $\hat{\mathbf{D}}$. These weights can be applied in Eq. (7) to obtain the room transfer function of the desired target directivity.

3. SOURCE DEVELOPMENT

The weighting vector $\mathbf{g}_T$ in Eq. (10) is calculated from the spherical harmonics transformed directivity pattern of the measurement source. The directivity pattern has to contain sufficiently large coefficients in every spherical harmonic order to generate valid weights.

The spherical harmonic coefficients of the directivity of any electro-acoustical source are determined by the size of the entire source and the aperture angle of the transducer in its enclosure [5]. This suggests the design of a new source for the synthesis method.

Eq. (10) projects spherical harmonic coefficients into weights for spatial orientations. This is an analogy to the discrete spherical harmonic expansion. Several spatial sampling strategies have been introduced to efficiently perform this operation [9]. The physical orientations of the source should be selected according to one of these strategies.

3.1. Simulation Model

A suitable measurement source can virtually be of any shape. Here, the shape is chosen to be spherical, allowing for rapid prototyping applying an analytical simulation model. In this model a transducer on a sphere is simplified as a radially vibrating cap [5]. It has to be noted that this model does not take into account partial modes or the physical interaction of transducers in a common volume. Based on the model it is possible to successively calculate the *squared aperture magnitude* and the radiated sound pressure.

3.2. Aperture Magnitude

The *aperture function*

$$a(\vartheta,\varphi) = a(\vartheta) = 1 - \varepsilon\left(\vartheta - \alpha/2\right) \quad (11)$$

of a single membrane on the north pole of a sphere spanning the aperture angle α can be formulated as a continuous function on the sphere. $\varepsilon(x)$ is the *Heaviside function* (or *unit step function*). The spherical harmonic coefficients of this function can be expressed as [10]

$$\hat{a}_n^m = \begin{cases} \sqrt{\pi(2n+1)}\int_{\cos(\alpha/2)}^{1} P_n(x)\,\mathrm{d}x & \text{if } m = 0 \\ 0 & \text{otherwise} \end{cases}. \quad (12)$$

$P_n(x)$ is the *Legendre polynomial* of the order n. The aperture can be rotated to any orientation applying the *Wigner-D* rotation [8].

$$E_{\mathrm{a}}(n) = \sum_{m=-n}^{n} |\hat{a}_n^m|^2 \quad (13)$$

describes the frequency independent squared magnitude of the aperture per spherical harmonic order n created by a specific aperture [11].

The squared aperture magnitude features a distinct maximum and several minima due to the *Legendre polynomial* $P_n(x)$. Reducing the aperture angle shifts the extremes towards higher orders n while simultaneously decreasing the absolute squared magnitude.

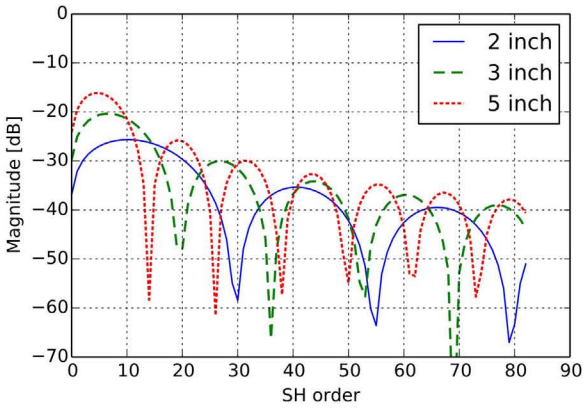


Figure 1: *Squared magnitude of the aperture functions of three transducer types on a sphere with radius* $r = 0.2\,\mathrm{m}$.

The radiation of a spherical harmonic order depends on the frequency and the size of the source (see 3.3). Thus, the squared magnitude depicted in Figure 1 describes the theoretical squared magnitude of each aperture, which is also a subject to an order-dependent low pass during the radiation. The squared magnitude in orders above 10 for the 5 inch transducer has to be neglected due to its frequency range, motivating the use of the 3 inch transducer.

3.3. Spherical Radiation

With the speed of sound c and the membrane displacement ξ, the aperture coefficient vector $\hat{\mathbf{a}}$ can be converted into a surface velocity coefficient vector [10]

$$\hat{\mathbf{v}}_{\mathrm{sphere}} = \mathrm{j}\,kc \cdot \xi \cdot \hat{\mathbf{a}}. \tag{14}$$

The *wave number* k introduces a frequency dependence. The surface velocity $\hat{\mathbf{v}}_{\mathrm{sphere}}$ can be transformed into the radiated sound pressure

$$\hat{\mathbf{p}}(r_{\mathrm{obs}}) = \rho_0 c^2 k \cdot \frac{h_n(k\,r_{\mathrm{obs}})}{h_n'(k\,r_{\mathrm{sphere}})} \cdot \xi \cdot \hat{\mathbf{a}} \tag{15}$$

at an arbitrary observation distance r_{obs} by multiplication with the acoustic impedance and the radial wave propagation [5]. The functions h_n and h_n' are the spherical *Hankel function* of the second kind in the n^{th} order and its derivative, respectively. Due to their properties, a sound source of a certain size can only radiate a limited range of orders at a particular frequency.

3.4. Source

The radius of the optimized source is $r = 0.2\,\mathrm{m}$. Due to the minima of the associated Legendre function in Eq. (11) a combination of three different transducer sizes (2, 3 and 5 inch) is used to generate a sufficiently large squared aperture magnitude in a wide range of spherical harmonic orders.

The 2 inch and 3 inch transducers are placed in a way that approximates a *Gaussian* sampling strategy of the order 11. The 5 inch transducers are placed accordingly in an order of 3. The physical locations are chosen to cover all required elevations for each transducer type. All azimuthal sampling points are reached by rotating the measurement source to 24 positions around the z-axis. A measurement for a source directivity of a spherical harmonic order of 11 generates 672 room impulses.

An additional tilt of the source to a second elevation and a rotation of the source to 48 positions in both elevations generates an approximation of a Gaussian sampling strategy of the order 23. This measurement generates 2688 room impulse responses.

The real behavior of the transducers such as membrane modes and deviations in the effective membrane area is expected to change the directivity of the real source compared to the simulated directivity.

3.5. Periphery

A turntable is used for the azimuthal rotation of the measurement source. To allow for an additional tilt the sphere is suspended on an axis piercing its eastern and western sides. The tilt is controlled by a step motor inside the sphere. The frame construction as seen in Figure 2 might have an impact on the directivity of the source.

Figure 2: *Optimized measurement source.*

4. DIRECTIVITY OF THE MEASUREMENT SOURCE

The calculation of the weighting vector $\mathbf{g}_T$ in Eq. (10) uses the directivity matrix $\hat{\mathbf{D}}$ in Eq. (8). The directivity of the transducers in a fixed orientation is measured in an anechoic chamber. The sound pressure at the measurement points is transformed into the spherical harmonic domain according to Eq. (5). The coefficient vectors for other source orientations are generated by multiplication with the Euler rotation term, as defined in 2.1.

It is crucial to choose a suitable sampling strategy for the measurement of the directivity. *Quadrature* samplings allow for a fast and exact spherical harmonic transformation of the sound pressure $\mathbf{p}$ at the sampling points into the directivity vector

$$\hat{\mathbf{d}} = \mathbf{Y}^H \cdot \mathrm{diag}(\mathbf{w}) \cdot \mathbf{p} \tag{16}$$

using the *quadrature weights* $\mathbf{w}$, as long as the spherical functions are order limited [12].

The measurement has to be done for a sufficiently high sampling order to prevent aliasing due to a violation of the required

order limitation. Different sampling strategies require different total numbers of spatial sampling points for a certain spherical harmonic order. The Gaussian quadrature sampling is easily usable and has a relatively high efficiency [9]. It is also quite robust against aliasing and useful for applications where rectangular samplings (a set of points at both constant azimuths and elevations) are beneficial [9].

All measurements are done using a Gaussian sampling strategy of the order $n = 82$. The elevation angle $\vartheta = 0°$ indicates an upward orientation in the spherical coordinate system, $\vartheta = 180°$ a downwards orientation, respectively.

The optimized source is placed on top of a turntable, allowing for a full $360°$ rotation in the azimuth angle φ. A swivel arm with a microphone is used to measure the sound pressure along one arc of sampling points down to $\vartheta = 90°$. This way, the upper hemisphere of the source radiation pattern is measured. Using the internal step motor to tilt the source by $180°$ allows for the subsequent measurement of the lower hemisphere. The directivity of each loudspeaker is measured separately, using an interleaved sweep measurement signal.

5. DEVIATION ANALYSIS

The directivity of the real source deviates from the simulated directivity simulated with the spherical cap model. For the synthesis method it is of interest to identify the deviation of the real directivity and to analyze its impact on the synthesis performance. For the directivity measurement it is necessary to gain knowledge about the radiated orders of the source to prevent aliasing effects in the directivity matrix $\hat{\mathbf{D}}$.

5.1. Directivity

Figure 3(a) exemplary shows the simulated and the respective measured directivity of a 5 inch transducer on the optimized measurement source at a frequency of 400 Hz. The transducer is orientated upwards at an angle of about 45 degrees. At this frequency the simulation and the measurement match quite well. An impact of the frame construction on the directivity cannot be identified.

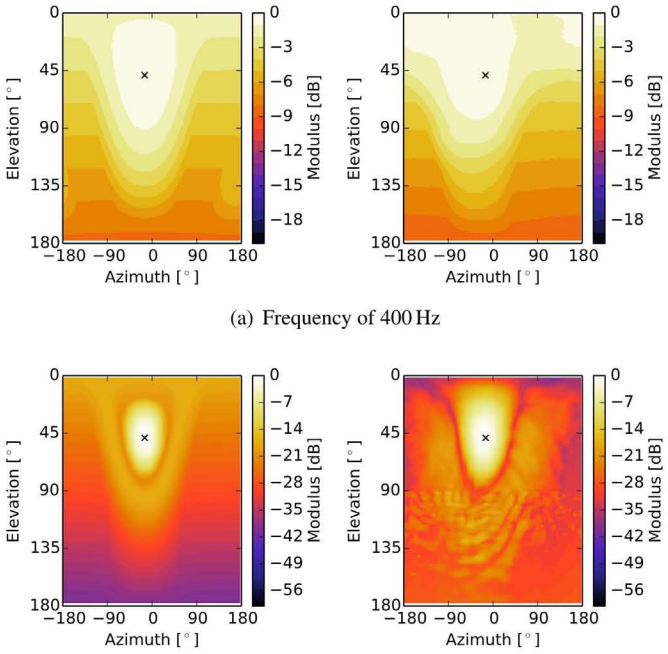
Figure 3(b) shows the same comparison at a frequency of 6400 Hz. Here, the measurement deviates clearly. Especially in angles below 90 degrees the radiation is changed. This effect can be explained with the measurement procedure. As described in section 4, the spherical body of the optimized source is tilted by $180°$ for the measurement of the lower hemisphere. In this position, the radiation of the regarded transducer is obstructed by the frame construction. Thus, the directivity can only be considered as measured correctly for the upper hemisphere and the measurement method introduces an additional error to the computation of the superposition weights in Eq. (10).

It is of interest to analyze the similarity of the simulated and measured directivity over the whole frequency range. The similarity of two spherical functions can be quantified using the spherical correlation [13]

$$C(f,g) = \oint_{S^2} \overline{f(\vartheta,\varphi)}\, g(\vartheta,\varphi)\, \mathrm{d}\Omega. \tag{17}$$

The continuous integral can be expressed by a weighted summation of the quadrature sampling points as normalized correlation

$$\tilde{C} = \frac{\sum_{i=1}^{N} \overline{f(\vartheta_i,\varphi_i)} g(\vartheta_i,\varphi_i) w_i}{\sqrt{E_f E_g}}, \tag{18}$$



(a) Frequency of 400 Hz

(b) Frequency of 6400 Hz

Figure 3: *Simulated (left) and measured (right) directivity of a 5 inch transducer of the optimized measurement source. Center of membrane marked with a black cross.*

with the energy of the two spherical functions f and g sampled at N points, $E_x = \sum_{i=1}^{N} |x(\vartheta_i,\varphi_i)|^2 w_i$ and w_i being the weights for the Gaussian quadrature sampling.

The correlation between simulation and measurement for the same transducer as shown in Figure 3(a) and Figure 3(b) is depicted in Figure 4. The correlation is high for low frequencies, which confirms the general validity of the simplified analytic model. For frequencies above 4 kHz the correlation drops, confirming that a measured directivity should be used for the computations of the weights in Eq. (10).

Changes of the effective membrane area and membrane modes introduce effects which render the simplified consideration of a radially vibrating cap on a sphere invalid. Furthermore, the constructive frame of the measurement source cannot be considered acoustically transparent in these frequency ranges and changes the radiation, especially due to the explained measurement error induced by tilting the source.

5.2. Radiation

For the synthesis it is important that the measurement source is capable of radiating all required spherical harmonic orders for the computation of the synthesis weights as shown in Eq. (10). Aliasing effects during the directivity measurement would additionally distort the directivity matrix $\hat{\mathbf{D}}$. Therefore it is also important to know about the maximum spherical harmonic order radiated by the measurement source.

The radiation can be simulated with the spherical cap model. Eq. (15) yields the radiation of a transducer on a spherical sound source in spherical harmonics. To look at the radiation of more

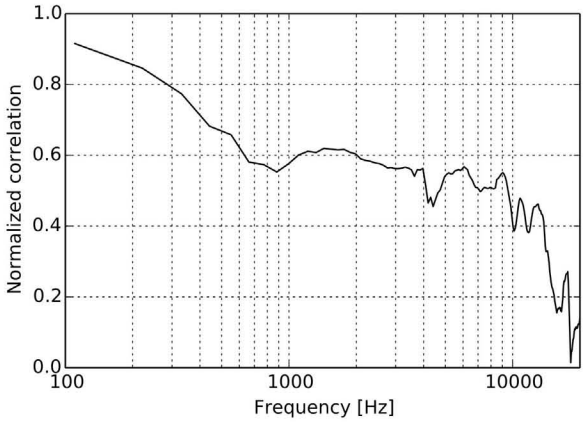
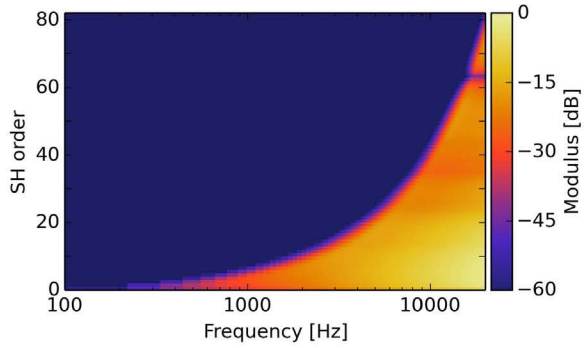
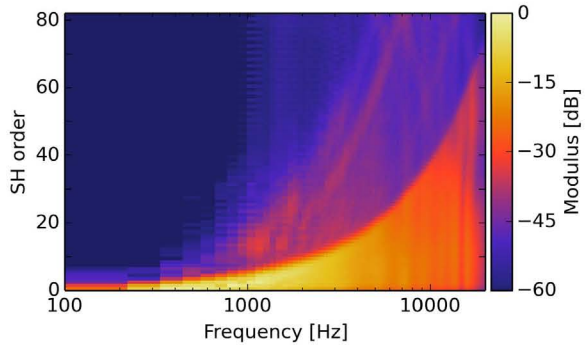


Figure 4: *Correlation between the simulation and measurement of a 5 inch transducer on the optimized measurement source.*



(a) Simulation result, maximum values.



(b) Measurement result, maximum values.

Figure 5: *Radiation in spherical harmonic orders, limited to a dynamic range of 60 dB.*

than one transducer, the magnitude of the coefficients of the radiation patterns can be summed up. Figure 5(a) shows the sum of the maximum magnitudes of each transducer in each order over the frequency. Since the absolute magnitude is of no interest, the data is normalized to the global maximum. The figure indicates the spherical harmonic orders that are radiated by the source. By rotating the source, the magnitude can be shifted to all coefficients within the respective order. The figure also indicates which orders have to be expected during the directivity measurement.

Orders up to 80 at a frequency of 20 kHz are expected to be excited by the optimized measurement source. In certain orders the radiation is minimal, coinciding with the effects explained in 3.2.The order limitation at each frequency is determined by the size of the source, as explained in 3.3. A sampling strategy with an order of 82 should be sufficient to prevent aliasing during the directivity measurement.

Figure 5(b) depicts the equivalent to Figure 5(a) for the measured directivity of the optimized source. The measurement has been done for one physical source orientation with a Gaussian sampling strategy of an order of 82. The data is normalized its global maximum.

A maximum order of 40 is radiated by the real source up to a frequency of 12 kHz. With the rotations to the measurement positions described in 3.4 this allows for the planned synthesis up to an order of 23.

The slope of the maximum sum of the measurement is not as steep as in the simulation. A parallel slope of higher order and lower magnitude can be observed, indicating a virtual source of a larger radius. This effect is most probably caused by reflections at the frame construction which lead to a virtual enlargement of the source. Aliasing effects can be expected starting at 5 kHz.

6. CONCLUSIONS

In this article the design of a specialized measurement source for room acoustical measurements with arbitrarily given source directivity is addressed. The parameters of the source geometry are defined using an analytical model of a vibrating spherical cap in a perfect spherical housing. The source is designed to be mounted on a computerized turntable enhancing the spatial resolution with a sequential measurement of different azimuthal orientations. Furthermore, the measurement source can be rotated around a horizontal axis, to increase the resolution in the elevation.

High-resolution directivity measurements of the actual prototype are used to confirm the results of the analytic calculation. The same measurements are used as a more realistic directivity $\hat{\mathbf{D}}$ for the synthesis of room impulse responses of arbitrary directivity patterns.

The device is expected to enhance the auralization of rooms for directive sound sources. Since a superposition approach is used, the desired directivity pattern can be synthesized in a post processing procedure. Even if room impulse responses with omnidirectional sources in accordance to ISO 3382 are desired, the presented measurement procedure can be used to enhance the result by synthesizing an even more omnidirectional sound source up to higher frequencies.

The prototype of the measurement source shows its capability of synthesizing directivity patterns up to a spherical harmonic order of about 23 for frequencies up to 12 kHz. This suggests great potential to improve auralizations of directive sound sources in measured acoustic environments. The measured directivity of the source is used for the computation of the synthesis weights. It has to be taken into account that for frequencies above 5 kHz this directivity contains errors due to the current frame construction and measurement procedure.

7. ACKNOWLEDGMENTS

The authors would like to thank the electrical workshop and the mechanical workshop of the ITA for their support during the construction of the measurement system as well as Ander Gaspar Perez Palacios for his commitment during the measurements and the post processing.

8. REFERENCES

[1] ISO 3382, *Acoustics – Measurement of room acoustic parameters – Part 1:Performance spaces*, ISO TC 43/SC 2, 2009.

[2] Martin Pollow, Pascal Dietrich, Martin Kunkemöller, and Michael Vorländer, "Synthesis of room impulse responses for arbitrary source directivities using spherical harmonic decomposition," in *Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, 2011.

[3] Martin Kunkemöller, Pascal Dietrich, and Martin Pollow, "Synthesis of room impulse responses for variable source characteristics," *Acta Polytechnica, Journal of advanced engineering, Prag*, vol. 51, pp. 69–74, 2011.

[4] Johannes Klein, Pascal Dietrich, Martin Pollow, and Michael Vorländer, "Optimized measurement system for the synthesis of transfer functions of variable sound source directivities for acoustical measurements," in *DAGA 2012*, 2012, pp. 345–346.

[5] Earl G. Williams, *Fourier Acoustics. Sound Radiation and Nearfield Acoustical Holography*, Academic Press, 1999.

[6] Martin Pollow, Johannes Klein, Pascal Dietrich, Gottfried K. Behler, and Michael Vorländer, "Optimized spherical sound source for room reflection analysis," in *Proceedings of the International Workshop on Acoustic Signal Enhancement (IWAENC)*, 2012.

[7] Milton Abramowitz and Irene A. Stegun, *Handbook of Mathematical Functions*, Dover Publications, 1970.

[8] Peter J. Kostelec and Daniel N. Rockmore, "FFTs on the Rotation Group," *Journal of Fourier Analysis and Applications*, vol. 14, pp. 145–179, 2008.

[9] Franz Zotter, "Sampling strategies for acoustic holography/holophony on the sphere," in *Proceedings of the NAG/DAGA 2009*, 2009.

[10] Martin Pollow and Gottfried K. Behler, "Variable directivity for platonic sound sources based on spherical harmonics optimization," *Acta Acustica United with Acustica*, vol. 95, pp. 1082–1092, 2009.

[11] Ilja N. Bronštein and Konstantin A. Semendjaev, *Handbook of Mathematics*, Springer, 2007.

[12] Franz Zotter, *Analysis and Synthesis of Sound Radiation with Spherical Arrays*, Ph.D. thesis, University of Music and Performing Arts Graz, 2009.

[13] Martin Pollow, Khoa-Van Nguyen, Olivier Warusfel, Thibaut Carpentier, Markus Müller-Trapet, Michael Vorländer, and Markus Noisternig, "Calculation of head-related transfer functions for arbitrary field points using spherical harmonics decomposition," *Acta acustica united with Acustica*, vol. 98, no. 1, pp. 72–82, 2012.

[14] Ilan Ben Hagai, Martin Pollow, Michael Vorländer, and Boaz Rafaely, "Acoustic centering of sources measured by surrounding spherical microphone arrays," *The Journal of the Acoustical Society of America*, vol. 130, pp. 2003–2015, 2011.

ASSESSING THE AUTHENTICITY OF INDIVIDUAL DYNAMIC BINAURAL SYNTHESIS

Fabian Brinkmann, Alexander Lindau, Martina Vrhovnik, Stefan Weinzierl

Audio communication group,
Technical University of Berlin, TUB
Berlin, Germany
fabian.brinkmann@tu-berlin.de

ABSTRACT

Binaural technology allows to capture sound fields by recording the sound pressure arriving at the listener's ear canal entrances. If these signals are reconstructed for the same listener the simulation should be indistinguishable from the corresponding real sound field. A simulation fulfilling this premise could be termed as perceptually *authentic*.

Authenticity has been assessed previously for static binaural resynthesis of sound sources in anechoic environments, i.e. for HRTF-based simulations not accounting for head movements of the listeners. Results indicated that simulations were still discernable from real sound fields, at least, if critical audio material was used.

However, for *dynamic* binaural synthesis to our knowledge – and probably because this technology is even more demanding – no such study has been conducted so far. Thus, having developed a state-of-the-art system for individual dynamic auralization of anechoic and reverberant acoustical environments, we assessed its perceptual authenticity by letting subjects directly compare binaural simulations and real sound fields. To this end, individual binaural room impulses were acquired for two different source positions in a medium-sized recording studio, as well as individual headphone transfer functions. Listening tests were conducted for two different audio contents applying a most sensitive ABX test paradigm. Results showed that for speech signals many of the subjects failed to reliably detect the simulation. For pink noise pulses, however, all subjects could distinguish the simulation from reality. Results further provided evidence for future improvements.

1. INTRODUCTION

As overall criteria for the quality of virtual acoustic environments, the perceived *plausibility* and *authenticity* has been proposed [1], [2]. Whereas the plausibility of a simulation refers to the degree of agreement with the listener's expectation towards a corresponding real event (agreement with an inner reference), authenticity refers to the perceptual identity with an explicitly presented real event (agreement with an external reference). While a non-individual data-based dynamic binaural synthesis has already been shown to provide plausible simulations [3], a dynamic synthesis based on individual binaural recordings appears to be a particularly promising candidate for a perceptually authentic acoustical simulation. Further, a formal assessment of the authenticity of state-of-the-art binaural technology would be of great practical relevance: Since nearly all currently known approaches to sound field synthesis (such as wave field synthesis, or higher order ambisonics) can be transcoded into binaural signals, a perceptually authentic binaural reproduction would provide a convenient reference simulation required for the strict, reliable and comprehensive evaluation of a wide variety of simulation approaches and systems [4].

Three empirical studies were found to be concerned with the authenticity of binaural simulations. However, all three studies assessed static auralization, i.e., simulations not accounting for natural head movements of the listeners. In order to allow for a convenient comparability, statistical significance of the observed results was assessed based on exact Bernoulli test statistics, if not initially given.

Langendijk and Bronkhorst [5] assessed the authenticity of individual binaural reproduction for six sound sources distributed evenly around the listener. Binaural signals were reproduced utilizing small earphones placed 1 cm in front of the concha with only little influence on the sound field of external sources. Band limited white noise bursts (500 Hz–16 kHz) were presented in a four interval 2AFC (alternative forced choice) paradigm where each sequence of four noise bursts contained three identical and one 'oddball'-stimulus in either second or third position, that had to be detected by the subjects. Detection rates across subjects were slightly but significantly above chance ($p_{\text{correct}} = 0.53$, 6 subjects, $N_{\text{total}} = 1800$ trials).

Moore et al. [6] conducted a similar listening test. Subjects participated twice in the experiment, and were considered untrained in the first run and trained in the second. A frontal sound source was auralized using cross-talk canceled (transaural) reproduction of individual binaural recordings. When presenting click or noise stimuli to trained subjects detection rates were again slightly but significantly above chance ($p_{\text{corr. click}} = p_{\text{corr. noise}} = 0.594$, 8 subjects, $N_{\text{total}} = 192$). *Untrained subjects*, however, were not able to detect the binaural simulation reliably ($p_{\text{corr. click}} = 0.5$, $p_{\text{corr. noise}} = 0.54$, $p_{\text{corr.testable}} = 0.675$ @ $\alpha = 0.05$ with 95% power, Dunn-Sidák corrected for multiple testing). Moreover, when using a *synthetic vowel sound*, the simulation was indistinguishable for both trained and untrained subjects ($p_{\text{corr.observed}} = 0.48$, $p_{\text{corr.testable}}$ as mentioned above).

Masiero [7] tested authenticity in a 3AFC test paradigm utilizing 24 sound sources distributed evenly around the listeners. Individual binaural signals were presented to 40 subjects through circumaural open headphones using noise, speech and music stimuli. Average detection rates were $p_{\text{corr. noise}} = 0.87$, $p_{\text{corr. speech}} = 0.74$, and $p_{\text{corr.music}} = 0.71$ (transformed to 2AFC detection rates for better comparability). While not being given originally by the authors, a *post hoc* inferential statistics analysis of the raw data revealed that for all three stimulus conditions detections rates were significantly above chance. Further, an

ANOVA conducted by Masiero showed the stimulus effect to be significant.

All three studies used some kind of head rest to control the subjects' head position. In addition, Moore et al. and Masiero monitored the subjects' head position with optic or magnetic tracking systems. Throughout his study, Masiero allowed for head movements between ±1°–2° rotation, and ±1–2 cm translation, respectively. Additionally, Masiero allowed his subjects to listen three times to the sequence of test stimuli whereas in the other two studies each condition was presented only once.

While – technically – being a far more demanding reproduction mode than static auralization, perceptual authenticity of *dynamic* binaural synthesis has not been assessed before. Moreover, a success of such an assessment has become more likely as number of technical improvements has been introduced recently: For example, new extraaural binaural headphones were presented *(BKsystem,* [8]) along with a perceptually optimized approach to the compensation of the headphone transfer function [9]. Further, an in-ear measurement systems for the reliable acquisition of individual binaural transfer functions (*PRECISE*, [9]) has been developed, and crossfade artifacts of dynamic binaural rendering have been minimized [10].

Further, as shown above, former studies achieved high statistical test power by cumulating test results over individuals and repeated trials while omitting a priori discussions of practical effect size and required test power. However, in order to limit the required methodological effort, and as individual performance was expected to be potentially quite different, we aimed at designing our test to produce practically meaningful results already on the level of individual subjects (cf. section 2.5).

2. METHOD

2.1. Setup

The listening tests were conducted in the recording studio of the *State Institute for Music Research*[1], Berlin ($V = 122\ \mathrm{m}^3$, $RT_{1\mathrm{kHz}} = 0.65$ s). Subjects were seated on a customized chair with an adjustable neck rest and a small table providing an armrest and space for placing the tactile interface used throughout the test (*Korg nanoKONTROL* Midi-Interface). An LCD screen was used as visual interface and placed 2 m in front of the subjects at eye level.

Two active near-field monitors (*Genelec 8030a*) were placed in front and to the right of the subjects at a distance of 3 m and a height of 1.56 m, corresponding to source positions of approximately 0° azimuth, 8° elevation (source 1) and -90° azimuth, 8° elevation (source 2). With a critical distance of 0.8 m and a loudspeaker directivity index of ca. 5 dB at 1 kHz, the source-receiver distance results in a slightly emphasized diffuse field component of the sound field. The height was adjusted so that the direct sound path from source 1 to the listening position was not blocked by the LCD screen. The source positions were chosen to represent conditions with minimal and maximal interaural time and level difference at a neutral head orientation (see test setup, Fig. 1).

[1] Staatliches Institut für Musikforschung, http://www.sim.spk-berlin.de/

Figure 1: *Listening test environment and used setup.*

For binaural reproduction, low-noise DSP-driven amplifiers and extraaural headphones were used, which were designed to exhibit only minimal influence on sound fields arriving from external sources while providing full audio bandwidth (*BKsystem*, [8]). Headphones were worn during the entire listening test, i.e. also during the binaural measurements, this way allowing for instantaneous switching between binaural simulation and corresponding real sound field. The subjects' head position was controlled using head tracking with 6 degrees of freedom (x, y, z, azimuth [head-above-torso orientation], elevation, lateral flexion) with a precision of 0.001 cm and 0.003°, respectively (*Polhemus Patriot*). A long term test of eight hours showed no noticeable drift of the tracking system.

Individual binaural transfer functions were measured at the blocked ear canal using *Knowles FG-23329* miniature electret condenser microphones flush cast into conical silicone earmolds. The molds were available in three different sizes, providing a good fit and reliable positioning for a wide range of individuals [9]. Phase differences between left and right ear microphones did not exceed ±2° avoiding audible interaural phase distortion [11].

The experiment was monitored by the investigator from a separate room with talk-back connection to the test environment.

2.2. Reproduction of Binaural Signals

The presence of headphones influences the sound field at the listeners' ears. Having considered an additional filter for compensating this effect [12], Moore et al. [6] concluded that headphones should not be used for direct comparisons of simulation and reality and consequently used transaural sound reproduction for their listening tests on authenticity. In contrast, we argue that a test on authenticity is not compromised as long as wearing the headphones (a) would affect real sound field and simulation in an identical manner and (b) would not mask possible cues for discriminating between the two. Condition (a) will be fulfilled by wearing the headphones both during measurement and simulation. For assessing condition (b), binaural room impulse responses (BRIRs) were measured with and without three types of headphones (*BKsystem*, *STAX SRS 2050 II*, *AKG K-601*) using a source at 2 m distance, -45° azimuth and 0° elevation for head-above-torso orientations in the range of ±80° azimuth. For this purpose, the head and torso simulator FABIAN equipped with a computer controlled neck joint for high precision and automated control of the head-above-torso orientation was used [13]. The headphone's influence was analyzed based on differences in the magnitude responses, and with respect to deviations of interaural time and level differences (ITD, ILD). For the *BKsystem*, magni-

tude response differences (Fig. 2, top left) show an irregular pattern with differences between approx. ±7.5 dB.

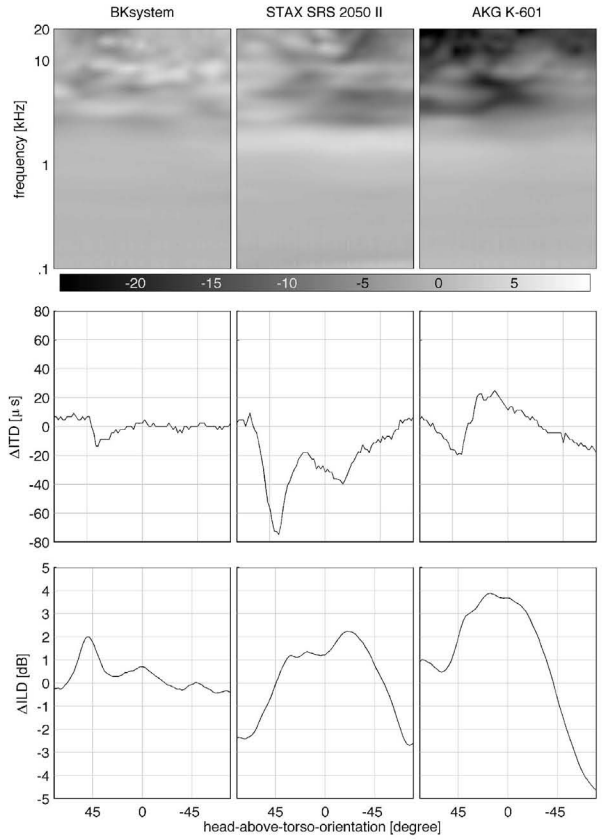


Figure 2: *Differences observed in BRIRs when measured with and without headphones for head-above-torso-orientations of between ±80° and for a source at -45° azimuth and 0° elevation. Top: Magnitude spectra (3rd octave smoothed, right ear, gray scale indicates difference in dB); Middle: ITDs; Bottom: ILDs.*

Whereas differences in magnitudes might influence localization in the median plane [14] the perceivable bandwidth of the signal remains largely unaffected making it unlikely that potential cues for a direct comparison would be eliminated. ITD and ILD differences are displayed in Fig. 2 (middle and bottom) and are believed to be inaudible for most head orientations. Assuming just audible differences of approximately 10-20 µs and 1 dB, respectively [1], only at 45°, where the ipsilateral ear is fully shadowed by the headphone, ILD differences slightly exceed the assumed threshold of audibility.

The observed differences are comparable to those found by Langendijk and Bronkhorst [5] who used small earphones near to the concha. Additionally, it is worth noting that differences were more than twice as high if conventional headphones were used (see Fig. 2).

2.3. Measurement of Individual Binaural Transfer Functions

Binaural room impulse responses and headphone transfer functions (HpTFs) were measured and processed for every subject prior to the listening test. *Matlab*® was used for audio playback, recording and processing the input signals. The head position of the subject was monitored using *Pure Data*. Communication between the programs was done by UDP messages. All audio processing was conducted at a sampling rate of 44.1 kHz.

Before starting the measurements, subjects put on the headphones and were familiarized with the procedure. Their current head position, given by azimuth and x/y/z coordinates was displayed on the LCD screen along with the target position given only by azimuth. Additionally, an acoustic guidance signal was played back through the headphones helping subjects finding the target azimuth for the subsequent measurement. The head tracker was calibrated with the test subject looking at a frontal reference position marked on the LCD screen. Subjects were instructed to keep their eye level aligned to the reference position during measurement and listening test, this way establishing also indirect control over their head elevation and roll. For training proper head-positioning, subjects were instructed to move their head to a specific azimuth and hold the position for 10 seconds. All subjects were quickly able to maintain a position with a precision of ±0.2° azimuth.

Then, subjects inserted the measurement microphones into their ear canals until they were flush with the bottom of the concha. Correct fit was inspected by the investigator. The measurement level was adjusted to be comfortable for the subjects while also avoiding limiting of both the DSP-driven loudspeakers and headphones.

BRIRs were measured for head-above-torso orientations between ±34° in azimuth and with a resolution of 2° providing smooth adaption to head movements [15]. The range was restricted to allow for a comfortable range of movements and convenient viewing of the LCD screen. Sine sweeps of an FFT order 18 were used for measuring transfer functions achieving a peak-to-tail signal-to-noise ratio (SNR) of approx. 80 dB for the BRIR at neutral head orientation without averaging [16].

The subjects started a measurement by pressing a button on the MIDI-interface after moving their head to the target position and reached it within ±0.1°. For the frontal head orientation, the target orientation had to be met also within 0.1 cm for the x/y/z-coordinates. For all other head orientations the translational positions naturally deviate from zero; in these cases subjects were instructed to meet the targeted azimuth only. During the measurement, head movements of more than 0.5° or 1 cm would have led to a repetition of the measurement, which rarely happened. These tolerance levels were set in order to avoid audible artifacts introduced by imperfect positioning [1][17].

Thereafter, ten individual HpTFs were measured per subject. To *a priori* account for potential positional variance in the transfer functions, subjects were instructed to move their head to the left and right in between individual headphone measurements. After all measurements, which took about 30 minutes, the investigator removed the microphones without changing the position of the headphones.

2.4. Post-Processing

In a first step, times-of-flight were removed from the BRIRs by means of onset detection and ITDs were calculated and stored separately. ITDs were reinserted in real time during the listening test, avoiding comb-filter effects occurring in dynamic auralization with non-time-aligned BRIRs and reducing the overall system latency [10]. Secondly, BRIRs were normalized with respect to their mean magnitude response between 200 Hz and 400 Hz.

Due to diffraction effects BRIRs exhibit an almost constant magnitude response in this frequency range making normalization especially robust against measurement errors and low-frequency noise. In a last step, BRIRs were truncated to 44100 samples with a squared sine fade out.

Individual HpTF compensation filters were designed using a weighted regularized least mean squares approach [18]. Filters of an FFT order 12 were calculated based on the average of ten HpTF per subject. Regularization was used to limit filter gains if perceptually required, the used approach is shortly explained here: HpTFs typically show distinct notches at high frequencies which are most likely caused by anti-resonances of the pinna cavities [19]. The exact frequency and depth of these notches strongly depends on the current fit of the headphones. Already a slight change in position might considerably detune a notch, potentially leading to ringing artifacts of the applied headphone filters [9]. Therefore, individual regularization functions were composed after manually fitting one or two parametric equalizers (PEQs) per ear to the most disturbing notches. The compensated headphones approached a *target band-pass* consisting of a 4th order Butterworth high-pass with a cut-off frequency of 59 Hz and a 2nd order Butterworth low-pass with a cut-off frequency of 16.4 kHz.

Finally, presentations of the real loudspeaker and the binaural simulation had to be matched to evoke equal loudness impressions. If assuming that signals obtained via individual binaural synthesis closely resemble those obtained from loudspeaker reproduction (cf. Fig. 3), loudness matching can be achieved by simply matching the RMS-level of simulation and real sound field. Hence, matching was pursued by adjusting the RMS-level of five second pink noise samples recorded from loudspeakers and headphones while the subject's head was in the frontal reference position. To account for the actual acoustic reproduction paths in the listening test, prior to loudness-matching, the headphone recordings were convolved with the frontal incidence BRIRs and the headphone compensation filter whereas the loudspeaker recordings were convolved with the target band-pass.

2.5. Test Design

The ABX test paradigm as part of the N-AFC test family provides an objective, criterion-free and particularly sensitive test for the detection of small differences [20], and thus seems appropriate also for a test on the authenticity of virtual environments. ABX-testing involves presenting a test stimulus (A), a hidden reference stimulus (B) and an open reference stimulus (X). Subjects may either succeed (correct answer) or fail (incorrect answer) to identify the test stimulus. Being a Bernoulli experiment with a (2AFC) guessing rate of 50%, the binomial distribution allows the calculation of exact probabilities for observed detection rates enabling tests on statistical significance.

If ABX tests are used to prove the authenticity of simulations, one should be aware that this corresponds to proving the null hypothesis H_0 (i.e., proving equality of test conditions). Strictly speaking, this proof cannot be given by inferential statistics. Instead, the approach commonly pursued is to establish empirical evidence that *strongly supports* the H_0, e.g. by rejecting an alternative hypothesis H_1 stating an effect of irrelevant size, e.g. a minimal increase of the empirical detection rate above the guessing rate (i.e., negating a minimum-effect hypothesis [21]).

When testing a difference hypothesis H_1, two kinds of errors can be made in the final decision: The type 1 (alpha) error refers to the probability of wrongly concluding that there was an audible difference although there was none. The type 2 (beta) error is made, if wrongly concluding that there was no audible difference although indeed there was one. The test procedure (i.e. the number of AFC decisions requested) is usually designed to achieve small type 1 error levels (e.g. 0.05), making it difficult (especially for smaller differences) to produce significant test results. If we aim, however, at proving the H_0 such a design may unfairly favor our implicit interest ('progressive testing'). In order to design a fair test we first decided about a practically meaningful effect size to be rejected and then aimed at balancing both error levels in order to statistically substantiate both the rejection and the acceptance of the null hypothesis, i.e. the conclusion of authenticity.

For the current listening test, a number of 24 trials was chosen per subject and for each test condition (i.e., one combination of source direction and stimulus type), ensuring that for 18 or more correct answers, the H_0 ($p_{\text{corr.}} = 0.5$) can be rejected, while for less than 18 correct answers, a specific H_1 of $p_{\text{corr.}} = 0.9$ can be rejected for one test condition, both at equal (i.e., fair) type 1 and type 2 error levels. The chosen statistical design also accounted for the fact that each subject had to conduct 4 repeated tests (i.e. error levels of 5% for individual tests were established by suitable Bonferroni correction). The rather high detection rate of $p_{\text{corr.}} = 0.9$ chosen to be rejected corresponds to our expectation that even small differences would lead to high detection rates, considering the very sensitive test design and the trained subjects available.

2.6. Test Procedure

Nine subjects with an average age of 30 years (6 male, 3 female) participated in the listening test, 3 of them were fairly and 6 of them highly experienced with dynamic binaural synthesis. No hearing anomalies were reported and all subjects had musical background (average 13 years of education). They could thus be regarded as expert listeners.

During the listening test three buttons (A/B/X) were displayed on the screen. Audio playback started, if the one of the buttons on the MIDI interface was pressed. To give the answer "A equals X", the corresponding button had to be pressed and held for a short time. Subjects could take their time at will and repeatedly listen to A, B and X before answering, controlling all interaction with the tactile MIDI interface.

Two audio contents were used: a pulsed pink noise (0.75 s noise, 1 s silence, 20 ms ramps) and an anechoic male speech recording (5 s). The latter was chosen as a familiar 'real-life' stimulus, while noise pulses were believed to best reveal potential flaws in the simulation. Further, the bandwidth of the stimuli was restricted using a 100 Hz high-pass to eliminate the influence of low frequency background noise on the binaural transfer functions. As mentioned already, four ABX tests were conducted per subject (2 sources x 2 contents) each consisting of 24 trials. The presentation order of content and source was randomized and balanced across subjects. On average, the test took about 45 minutes. To avoid a drift in head position, subjects were instructed to move their head back to the reference position once between each trial and to keep the head's orientation at approx. 0° elevation throughout the test.

Dynamic auralization was realized using the fast convolution engine *fWonder* [13] in conjunction with an algorithm for real-time reinsertion of the ITD [10]. *fWonder* was also used for

applying (a) the HpTF compensation filter and (b) the loudspeaker target band-pass. The playback level for the listening test was set to 60 dB(A). BRIRs used in the convolution process were dynamically exchanged according to the subjects' current head-above-torso orientation, and playback was automatically muted if the subject's head orientation exceeded 35° azimuth.

2.7. Physical Verification

Prior to the listening test, acoustic differences between test conditions were estimated based on measurements with the FABIAN dummy head. Therefore, FABIAN was placed on the chair and BRIRs and HPTFs were measured and post-processed as described above. In a second step, BRIRs were measured as being reproduced by the headphones and the simulation engine described above. Differences between simulation and real sound field for the left ear and source 1 are depicted in Fig. 3.

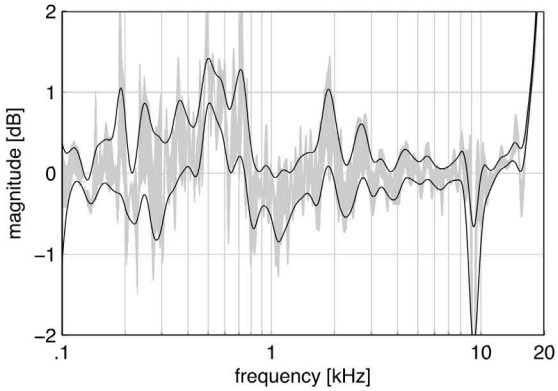


Figure 3: *Differences between binaural simulation and real sound field for source 1 and left ear. The grey area encloses the range of differences observed for all head-above-torso orientations between ±34°. For ease of interpretation, the range of differences is shown again after applying 6th octave smoothing (black lines).*

At a notch frequency in the HpTF at 10 kHz, differences reached up to 6 dB. However, this was assumed to be perceptually irrelevant since the bandwidth of the notch was less than a 10th octave. Above 3 kHz differences were in a range of ±0.5 dB. Somewhat larger and presumably audible deviations of up to ±2 dB were observed between 100 Hz and 3 kHz which were potentially caused by time variance of electro-acoustic transducers. Altogether, Fig. 3 shows comparable error patterns as Fig. 7b in Moore et al. [6].

3. RESULTS

Results of the ABX listening test are summarized in Fig. 4 for all subjects. A clear difference in detection performance was found between contents: While for the pulsed noise subjects were able to discriminate simulation and real sound field (all individual tests were statistically significant, see sect 2.5. for the description of the statistical test), for the speech stimulus about half of them were not (55% significant tests). This increased uncertainty is also reflected in larger variance across subjects. Moreover, a tendency for higher detection rates ($p_{\text{corr.}}$) was found for source 2 ('s2') compared to source 1 ('s1'). Although statistical analysis of detectability was conducted on the level of individual subjects, observed average detection rates are given for better comparability to earlier studies: $p_{\text{corr. noise s1}} = 0.978$, $p_{\text{corr. noise s2}} = 0.991$, $p_{\text{corr. speech s1}} = 0.755$, and $p_{\text{corr. speech s2}} = 0.829$.

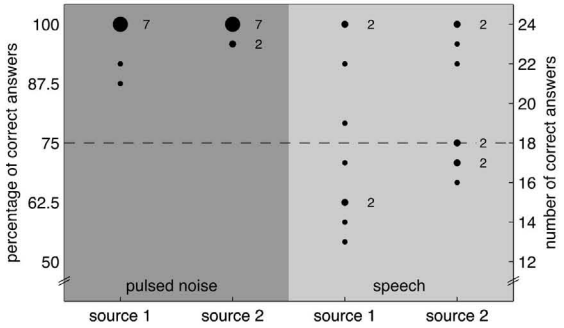


Figure 4: *Listening test results of nine subjects and for each test condition. Dots indicate percentage/number of correct answers for each tested condition; singular numbers indicate subjects with identical detection results. Dots on or above the dashed line indicate statistically significant differences.*

Differences between stimuli could also be found when comparing the average duration needed for making decisions (significantly higher for speech: 38 s vs. 15 s, $p < 0.01$, Wilcoxon signed rank test for dependent samples). Furthermore, increased head movements were found for speech (interquartile range 20° vs. 8° azim., $p < 0.01$, Wilcoxon signed rank test for dependent samples), indicating an extended search behavior adopted by subjects.

During auralization, BRIRs were selected solely based on the subjects' head-above-torso orientation. Hence, unobserved differences in the remaining degrees of freedom (x, y, z, elevation, lateral flexion) might have caused audible artifacts. Therefore, head tracker data were recorded and used for a post hoc analysis of deviations between head position during binaural measurements and ABX tests: For x, y, z coordinates, deviations were found to have been smaller than 1 cm for 95% of the time and never exceed 2 cm which is well within limits given by Hiekkanen et al. [17]. Differences in head elevation (tilt) and in lateral flection (roll) rarely exceeded 10° and were below 5° for 90% of the time. This may have caused audible artifacts occasionally [1], but a systematic influence on the results is unlikely.

When asked for the qualities of perceived differences between simulation and reality after the listening test, subjects named coloration (7x), slight differences in loudness (2x), and spaciousness (1x). Furthermore, two subjects reported a hissing or resonating sound in the decay of the noise pulses.

4. DISSCUSSION AND OUTLOOK

In the present study we assessed whether a state-of-the-art individual dynamic binaural simulation of an echoic environment can still be discriminated from the corresponding real sound field (test of 'perceptual authenticity'). To this end, measurement and post-processing of individual binaural transfer functions was demonstrated to be feasible within a reasonable amount of time, while obtaining a sufficient SNR and avoiding excessive test subject fatigue. Further, listening tests were conducted immedi-

ately after the measurements (i.e., – due to the minimization of deviations caused by time variability – resembling a best case scenario when aiming at proving authenticity) using a sensitive ABX test paradigm.

In accordance with earlier studies, we found that for a pulsed pink noise sample all subjects could reliably detect a difference between reality and simulation (individual detection rates between 87.5% and 100%). In case of the speech sample, however, only about half of the subjects still perceived a difference (individual detection rates between 54% and 100%). The higher detectability for the noise stimulus can be explained by its broadband and steady nature, supporting the detection of coloration, which, according to the subjects, was perceived as the major difference. Further, in considering this, also the mentioned loudness differences might be related to remaining spectral deviations.

Furthermore, higher detection rates were observed for source 2 as compared to source 1. These could be explained by occasionally observed slight discontinuities in the extracted ITD, most probably due to lower SNR at the contralateral ear. Additionally, low SNR might have led to larger measurement errors potentially perceivable as coloration.

Further, a tendency for interaction between source and type of stimulus was observed, as across all subjects, detection rate was by far lowest for source 1 and the speech stimulus ($p_{corr.s1.noise}$ = 75.5%). The observed value indicates that for this condition the group's detection performance was at threshold level (discrimination between simulation and reality in 50% of the cases, equalling 75% in a 2AFC paradigm).

On overall, the observed detection rates were higher than those reported in previous studies, although the precision of the binaural reproduction was comparable [6]. Hereby, our test design allowing subjects to switch at will between stimuli before making final decisions, may be assumed to be much more sensitive to small flaws of the simulation than sequence-based presentations applied in previous studies. This is also indicated by the fact that six subjects reported to have felt to be merely guessing although four of them produced significant detection results for one source of the speech stimulus. In addition, results indicate that it is still more demanding to realize an authentic interactive real time simulation as compared to static auralization. This was somehow expectable as extended abilities of a simulation naturally go together with extended potential for perceptual issues (e.g., with respect to crossfading, latency, or spatial discretization).

Moreover, and in contrast to former studies, our test included simulating a reverberant environment. Future tests which are planned to be conducted in an anechoic chamber and a concert hall will reveal whether the simulation of reverberant environments resembles a specific challenge.

The 'hissing' sound perceived by two subjects might be an artefact related to slightly mistuned headphone filters, indicating the potential for future improvements of our simulation as e.g. with respect to perceptually more robust headphone filter design. Further, an optimization of individual ITD modelling appears advisable and will be pursued in the future.

5. SUMMARY

A test of authenticity was conducted for the first time for a dynamic individual binaural simulation. Results showed that when by applying a sensitive test design the simulation was always clearly distinguishable from the real sound field, at least for critical sound source positions and if presenting noise bursts. However, for male speech, resembling a typical 'real-life' audio content and for a non-critical source position, half the subjects failed to reliably discriminate between simulation and reality, and averaged across subjects performed at threshold level.

6. ACKNOWLEDGMENTS

This work was funded by the German Research Foundation (DFG WE 4057/3-1).

7. REFERENCES

[1] J. Blauert, *Spatial Hearing. The psychophysics of human sound localization*, MIT Press, Revised Edition, Massachusetts, USA, 1997.

[2] R. S. Pellegrini, "A virtual reference listening room as an application of auditory virtual environments," Ph.D. thesis, University Bochum, 2001.

[3] A. Lindau and S. Weinzierl, "Assessing the plausibility of virtual acoustic environments," *Acta Acust. united Ac.*, vol. 98, no. 5, pp. 804-810, 2012.

[4] H. Wierstorf, A. Raake, M. Geier and S. Spors, "Perception of focused sources in wave field synthesis," *J. Audio Eng. Soc.*, vol. 61, no. 1/2, pp. 5-16, 2013.

[5] E. H. A. Langendijk and A. W. Bronkhorst, "Fidelity of three-dimensional-sound reproduction using a virtual auditory display," *J. Acoust. Soc. Am.*, vol. 107, no. 1, pp. 528-537, 2000.

[6] A. H. Moore, A. I. Tew and R. Nicol, "An initial validation of individualized crosstalk cancellation filters for binaural perceptual experiments," *J. Audio Eng. Soc. (Engineering Report),* vol. 58, no. 1/2, pp. 36-45, 2010.

[7] B. Masiero, "Individualized Binaural Technology. Measurement, Equalization and Perceptual Evaluation," Ph.D. thesis, RWTH Aachen, 2012.

[8] V. Erbes, F. Schulz, A. Lindau and Stefan Weinzierl, "An extraaural headphone system for optimized binaural reproduction," *Fortschritte d. Akustik: Tagungsband d. 38. DAGA [German annual acoustic conference]*, pp. 313-314, Darmstadt, Germany, March, 2012.

[9] A. Lindau and F. Brinkmann, "Perceptual evaluation of headphone compensation in binaural synthesis based on non-individual recordings," *J. Audio Eng. Soc.*, vol. 60, no. 1/2, pp. 54-62, 2012.

[10] A. Lindau, J. Estrella and S. Weinzierl, "Individualization of dynamic binaural synthesis by real time manipulation of the ITD," in *Proc. 128th AES Convention*, London, UK, May 22-25, 2010.

[11] A. W. Mills, "On the minimum audible angle," J. Acoust. Soc. Am., vol. 30, no. 4, pp. 237-246, 1958.

[12] A. H. Moore, A. I. Tew and R. Nicol, "Headphone transparification: A novel method for investigating the externalisation of binaural sounds," in *Proc. 123rd AES Convention, Convention Paper 7166*, New York, USA, October, 2007.

[13] A. Lindau, T. Hohn and S. Weinzierl, "Binaural resynthesis for comparative studies of acoustical environments," in *Proc. 122th AES Convention, Convention Paper 7032*, Vienna, Austria, May, 2007.

[14] W. M. Hartmann and A. Wittenberg, "On the externalization of sound images," *J. Acoust. Soc. Am.*, vol. 99, no. 6, pp. 3678-3688, 1996.

[15] A. Lindau and S. Weinzierl, "On the spatial resolution of virtual acoustic environments for head movements on horizontal, vertical and lateral direction," in *Proc. EAA Symposium on Auralization*, Espoo, Finland, June 15-17, 2009.

[16] S. Müller and P. Massarani, "Transfer function measurement with Sweeps. Directors's cut including previously unreleased material and some corrections," *J. Audio Eng. Soc. (Original Release)*, vol. 49, no. 6, pp. 443-471, 2001.

[17] T. Hiekkanen, A. Mäkivirta and M. Karjalainen, "Virtualized listeneing tests for loudspeakers," *J. Audio Eng. Soc.*, vol. 57, no. 4, pp. 237-251, 2009.

[18] S. G. Norcross, M. Bouchard and G. A. Soulodre, "Inverse Filtering design using a minimal phase target function from regularization," in *Proc. 121th AES Convention, Convention Paper 6929*, San Francisco, USA, October 5-8, 2006.

[19] H. Takemoto, P. Mokhtari, H. Kato, R. Nishimura and K. Iida, "Mechanism for generating peaks and notches of head-related transfer functions in the median plane," *J. Acoust. Soc. Am.*, vol. 132, no. 6, pp. 3832-3841, 2012.

[20] L. Leventhal, "Type I and type 2 errors in the statistical analysis of listening tests," *J. Audio Eng. Soc.*, vol. 34, no. 6, pp. 437-453, 1986.

[21] K.R. Murphy and B. Myors, "Testing the Hypothesis That Treatments Have Negligible Effects: Minimum-Effect Tests in the General Linear Model," *J. Appl. Psychol.*, vol. 84, no. 2, pp. 234-248, 1999.

EFFICIENT PHANTOM SOURCE WIDENING AND DIFFUSENESS IN AMBISONICS

Franz Zotter, Matthias Frank

Institute of Electronic Music and Acoustics
Univ. of Music and Performing Arts Graz
Austria, {zotter, frank}@iem.at

Matthias Kronlachner,

Univ. of Music and Performing Arts Graz
Austria, m.kronlachner@gmail.com
http://matthiaskronlachner.com

Jung-Woo Choi,

Dept. of Mechanical Engineering, KAIST
Korea Advanced Institute of Science and Technology
khepera@kaist.ac.kr

ABSTRACT

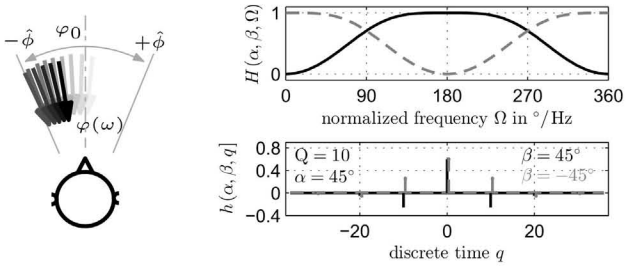
Object-based spatial audio considers virtual sound sources having a width/diffuseness parameter. This parameter aims at controlling the perceived width or diffuseness of the auditory object, or phantom source, created by the renderer. Width/diffuseness provides an important salience parameter that is independent of perceived direction and timbre. A highly efficient sparse filter structure for two-channel stereophony was described and tested recently, but it becomes ineffective for most parts of a large audience. This paper presents phantom source width/diffuseness control for Ambisonics. The new approach is a remarkably elegant application of the previously described stereo phantom source widening on Ambisonics. Compared with former experimental data, our experiments show a greater freedom of increasing the width and widening that works for a larger listening area.

1. INTRODUCTION

Special filter structures for increasing the width of two-channel stereophonic phantom sources have been discussed since the nineteen fifties, e.g. [1, 2, 3]. More general frequency domain fast block filter all-pass designs with random phase were most successful for phantom source widening, [4, 5, 6], until recently a special sparse FIR phase-based and amplitude-based filter structure could be proposed [7, 8]. However, experiment and practice show that two-channel stereo widening or diffuseness degrades for more than 30cm laterally shifted listening positions. Moreover, widening/diffuseness does not exceed the loudspeaker pair.

For phantom source widening on many loudspeakers, vector-base amplitude panning (VBAP) using a frequency-dependent panning angle was presented in [9]. The effectiveness of the approach could be heard at audio engineering workshops[1]. The practical implementation employs fast block filters that represent the frequency-dependent panning. However, the steep stop band transitions of these filters aren't continuously differentiable, hence careful FIR approximation becomes necessary.

Recent work [10] lets us expect alternative, simple filters in Ambisonics, but we currently lack suitable algorithms. We therefore consider Ambisonic encoding of a plane wave, cf. Fig. 1(a),



(a) Basic idea. (b) $H(\alpha, \beta, \Omega)$ has a sparse time response.

Figure 1: Sound with frequency-dispersed arrival direction (a) is perceived as widened. Can it be created by the sparse filter in (b)?

$$p(\omega, \boldsymbol{r}) = e^{\mathrm{i}k\,\boldsymbol{s}^{\mathrm{T}}\boldsymbol{r}} \tag{1}$$

whose direction of arrival (DOA) shall be frequency-dependent $\boldsymbol{s} = \boldsymbol{s}(\omega)$. k is the wave number ω/c, ω is 2π times frequency f, c is the speed of sound, $\boldsymbol{s}$ is a direction vector $\|\boldsymbol{s}\| = 1$, and $\boldsymbol{r}$ is a point of observation. Despite frequency dependency disperses the DOA, the sound pressure stays perfectly unitary $|p(\omega, \boldsymbol{r})| = 1$.

The Fourier transform pair of a phase modulated cosine [11] yields a sparse time response when used to define a time-invariant frequency response, cf. appendix, with the normalized frequency $\Omega = \omega\mathrm{T}$,

$$H(\alpha, \beta, \Omega) = \cos[\alpha\,\cos(\Omega) + \beta], \tag{2}$$

$$\xrightarrow{\mathcal{IFT}} h(\alpha, \beta, t) = \sum_{\lambda=-\infty}^{\infty} \cos(\tfrac{\pi}{2}\,|\lambda| + \beta)\, J_{|\lambda|}(\alpha)\, \delta(t + \lambda).$$

Its simplest discrete-time implementation $h(\alpha, \beta, q]$, cf. Fig. 1(b), uses the integer sample index q, and is nonzero only at $q = \lambda\mathrm{Q}$,

$$h(\alpha, \beta, \lambda\mathrm{Q}] = \cos(\tfrac{\pi}{2}|\lambda| + \beta)\, J_{|\lambda|}(\alpha), \tag{3}$$

with $\lambda, q \in \mathbb{Z}$, $\mathrm{Q} \in \mathbb{N}$. For small α the filter is truncated in $|\lambda|$ and implemented efficiently as sparse FIR filter. Such filters were employed in [8] for two-channel stereo phantom source widening.

[1] workshop by Ville Pulkki, "Time-frequency processing of spatial audio", presented at AES 128th Conv., London; AES 40th Int. Conf. on Spatial Audio, Tokyo; AES 129th Conv., San Francisco; EAA Winter School Cutting Edge in Spatial Audio, Merano 2013

Assuming a dispersion constant $\hat{\phi}$ and an average DOA φ_0, this filter might intuitively provide a frequency-dispersed DOA

$$\boldsymbol{s}(\Omega) = \begin{pmatrix} \cos[\hat{\phi}\cos(\Omega) + \varphi_0] \\ \sin[\hat{\phi}\cos(\Omega) + \varphi_0] \end{pmatrix}. \tag{4}$$

Based on this, an Ambisonic representation of a dispersive DOA encoding system is developed in section two. The third section describes a listening experiment to clarify whether the algorithm widens the phantom source at three different listening positions more effectively than a comparable two-channel stereo implementation. After discussing how good results of the perceptual study correlate with the $\mathrm{IACC_{E3}}$ measure, section four highlights generalization of the Ambisonic widening algorithm. Non-sinusoidal DOA dispersion curves might be favorable for Ambisonic orders higher than 3 to avoid DOA accumulation points. Moreover, to widen/diffuse entire Ambisonic productions, rotation matrices rather than encoders are defined. For 3D Ambisonics, spherical-cap-filling and isotropic widening is briefly sketched.

2. FREQUENCY-VARYING AMBISONIC ENCODER

The incoming field p is classically represented in 2D Ambisonics by the encoding coefficients b_m that drive the equation

$$p(\omega, r, \varphi) = \underbrace{\sum_{m=-\infty}^{\infty} \mathrm{i}^m \, J_m(kr) \, \Phi_m(\varphi) \;\; b_m(\varphi_\mathrm{s})}_{\text{represented by playback system}}, \tag{5}$$

with the orthonormal circular harmonics

$$\Phi_m(\varphi) = \sqrt{\tfrac{2-\delta_m}{2\pi}} \begin{cases} \cos(m\varphi), & \text{for } m \geq 0, \\ -\sin(m\varphi), & \text{for } m < 0. \end{cases} \tag{6}$$

A plane-wave field arriving from the direction of the polar angle φ_0 would just be encoded by a set of frequency-independent scalars $b_m(\varphi_0) = \Phi_m(\varphi_0)$. A DOA-dispersed plane wave around φ_0 requires encoding by $\Phi_m[\hat{\phi}\,\cos(\Omega) + \varphi_0]$ and hence responses are related to Eq. (2) by $\alpha = m\hat{\phi}$ and $\beta = m\varphi_0$ or $\beta = m\varphi_0 + \pi/2$,

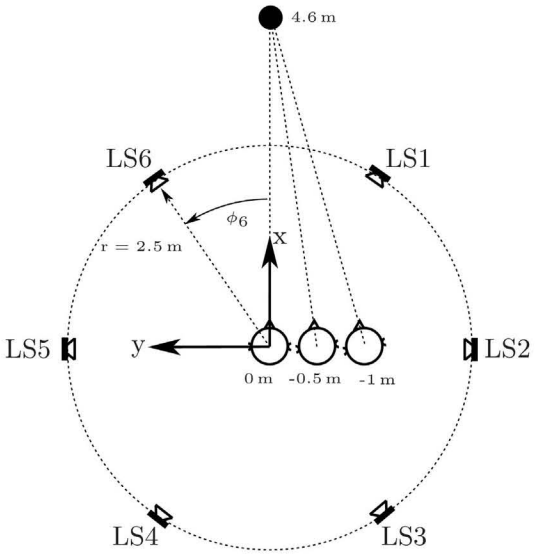
$$b_m(\Omega) = \sqrt{\tfrac{2-\delta_m}{2\pi}} \begin{cases} \cos[m\hat{\phi}\cos(\Omega) + m\varphi_0], & m \geq 0, \\ -\sin[m\hat{\phi}\cos(\Omega) + m\varphi_0], & m < 0. \end{cases}$$

Correspondingly, an encoding system is defined by the transform pair in Eq. (2) and permits efficient sparse FIR implementation using Eq. (3). Herewith, impulse responses $b_m[q]$ of Ambisonic encoding with dispersed DOA in the inverval $[\varphi_0 - \hat{\phi},\, \varphi_0 + \hat{\phi}]$ become nonzero only at $\lambda\mathrm{Q}$. For causal and finite responses, λ is shifted by a causality-bringing truncation offset $\Lambda \in \mathbb{N}$, and it is truncated to $0 \leq \lambda \leq 2\Lambda$. The $2\Lambda + 1$ nonzero entries of $b_m[q]$ are defined by Eq. (3) and lie in the discrete-time inverval $q \in [0,\, 2\Lambda\mathrm{Q}]$,

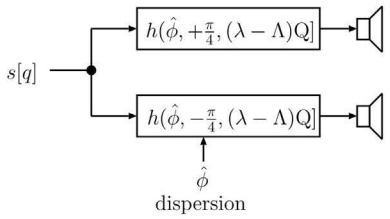
$$b_m[\lambda\mathrm{Q}] = \sqrt{\tfrac{2-\delta_m}{2\pi}} \begin{cases} h(m\hat{\phi}, m\varphi_0, (\lambda-\Lambda)\mathrm{Q}], & m \geq 0, \\ h(m\hat{\phi}, m\varphi_0 + \pi/2, (\lambda-\Lambda)\mathrm{Q}]. & m < 0. \end{cases} \tag{7}$$

Discrete-time Ambisonic signals $\chi_m[q]$ encoding a sound signal $s[q]$ are obtained by convolution with this single-input-multiple-output (SIMO) system of impulse responses

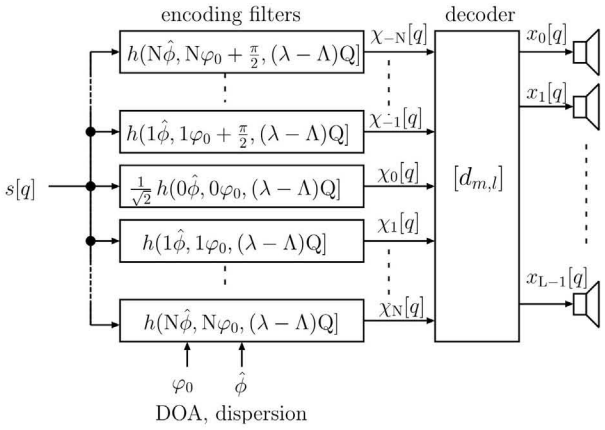
$$\chi_m[q] = s[q] \star b_m[q]. \tag{8}$$



(a) Setup of the listening experiment.



(b) Stereo widening around $\varphi_0 = 0$.



(c) Proposed Ambisonic widening around φ_0.

Figure 2: Experimental setup and signal processing schemes for testing widened perceived (phantom) sources with dispersion $\hat{\phi}$. Stereo widening is used to supply LS6 and LS1, and Ambisonic widening supplies LS1…LS6 and uses N=2.

The experimental evaluation that follows uses second-order Ambisonics, i.e., it restricts m to $|m| \leq \mathrm{N}$ with $\mathrm{N} = 2$. We chose the truncation offset $\Lambda = 9$ and a time grid of $\mathrm{Q} = 110$ samples between nonzero entries, i.e. $\mathrm{T} \approx 2.5\,\mathrm{ms}$ at the sample rate $f_\mathrm{s} = 44.1\,\mathrm{kHz}$.

3. EXPERIMENT

The listening experiment evaluates the perceived source extent of the Ambisonic widening and compares it to the stereo widening. The experiment hereby investigates the controllability of the widening effect at different listening positions.

3.1. Setup and Conditions

Fig. 2(a) shows the experimental setup consisting of a hexagonal ring of 6 Genelec 8020 loudspeakers at a radius of $\mathrm{r} = 2.5\,\mathrm{m}$. The hexagonal layout is suitable for second order Ambisonics. As it contains a loudspeaker pair at $\pm 30°$, comparison to a two-channel stereo widening algorithm is possible. The height of all loudspeakers was set to $1.2\,\mathrm{m}$ which is also the ear height of the subjects. The experiment was performed in the IEM CUBE, a $10.3\,\mathrm{m} \times 12\,\mathrm{m} \times 4.8\,\mathrm{m}$ large room with a mean reverberation time of $470\,\mathrm{ms}$ that fulfills the recommendation for surround reproduction in ITU-R BS.1116-1 [12]. The central listening position lies within the effective critical distance. The subjects were asked to fix a reference point at $4.6\,\mathrm{m}$ in the frontal direction at all listening positions.

For the regular hexagonal setup, an optimal $\max\text{-}\boldsymbol{r}_\mathrm{E}$ Ambisonic decoder is obtained by $d_{lm} = \cos[\frac{m\pi}{2(\mathrm{N}+1)}]\,\Phi_m[-\frac{\pi\,(1+2l)}{6}]$, cf. [13], to get six loudspeaker signals from five Ambisonic signals $\chi_m(t)$

$$x_l[q] = \sum_{m=-2}^{2} d_{lm}\;\chi_m[q], \quad \forall l = 0\ldots 5. \tag{9}$$

Dispersive encoding Eq. (7) is used with $\mathrm{Q} = 110$ and $\Lambda = 9$ ($0 \le \lambda \le 2\Lambda$). The DOA dispersion $\hat{\phi}$ is investigated for $\varphi_0 = 0$.

As a reference, an amplitude-based two-channel stereo widener was tested using filters derived from Eq. (3) to generate signals for loudspeaker one and six, aiming at widened frontal sound, cf. [8],

$$x_{6,1}[q] = s[q] \star h(\hat{\phi}, \pm\tfrac{\pi}{4}, (\lambda - \Lambda)\mathrm{Q}], \tag{10}$$

also using $\mathrm{Q} = 110$, $\Lambda = 9$ ($0 \le \lambda \le 2\Lambda$), and a varying $\hat{\phi}$.

The algorithms were fed with $22\,\mathrm{s}$ of English speech from EBU SQAM CD [14] and adjusted to 65dB(A). Speech was less sensitive to subtle perceivable differences that do not affect the perceived spatial extent but could indirectly distort the results.

Table 1: *Stereophonic and Ambisonic conditions of the listening experiment share cond. 1 and were adjusted to cover similar widths.*

cond.	stereo	Ambisonics
1	$\hat{\phi} = 0°$	$\hat{\phi} = 0°$
2	$\hat{\phi} = 30°$	-
3	$\hat{\phi} = 50°$	-
4	$\hat{\phi} = 70°$	-
5	$\hat{\phi} = 90°$	-
6	-	$\hat{\phi} = 20°$
7	-	$\hat{\phi} = 35°$
8	-	$\hat{\phi} = 47°$
9	-	$\hat{\phi} = 65°$

For both stereo and Ambisonic algorithms, 5 different conditions with varying dispersion $\hat{\phi}$ were tested, cf. Tab. 1. Dispersion values of stereo conditions were chosen to be comparable to known data on phase-based stereo widening [7, 8] that used the same input speech signal. Dispersion values of the Ambisonic conditions were adjusted in a preliminary test at the central listening position as to obtain a perceived extent similar to the extent of the stereophonic conditions. The zero-dispersion condition $\hat{\phi} = 0°$ produces exactly the same loudspeaker signals for stereo and Ambisonics on the given setup. Hence this condition was only tested once, cf. condition 1 in Tab. 1. Each subject completed 6 MUSHRA-like comparisons, each of which displaying the 9 conditions in random order. In this way, each subject gave a full comparison twice at each of the three listening positions.

In the MUSHRA-like comparison, subjects were allowed to seamlessly switch between the playback of the 9 conditions. The user interface was placed on their lap. In each comparison, subjects rated the perceived spatial extent of the 9 selectable stimuli on 9 corresponding quasi-continuous sliders from "narrow" to "wide" (the test used the German terms „schmal" and „breit"). To support subjects in organizing their ratings, they were provided a button to re-arrange the 9 selectable conditions by ascending slider values.

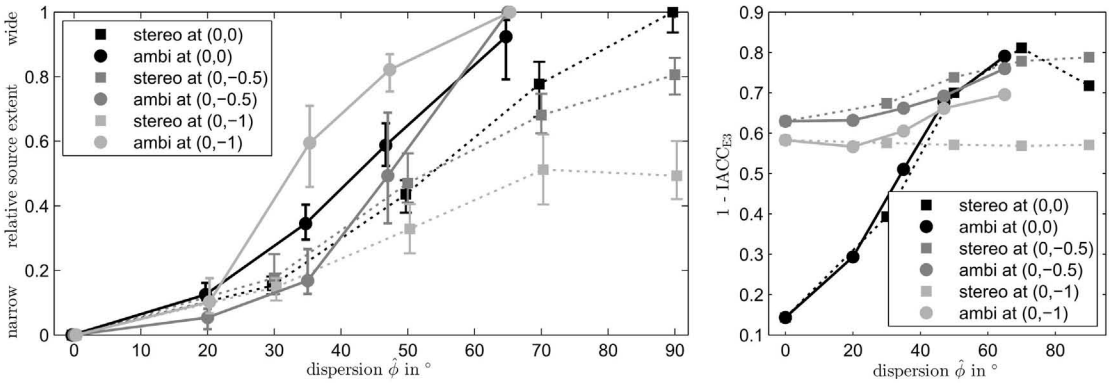
All of the 12 subjects (all male, age range: 27-39 years, median: 31 years) were members of the Institute of Electronic Music and Acoustics and familiar with spatial audio.

3.2. Results

Within each subject, the repetitions of the same comparison task correlated at least with 73%. The mean correlation was 92%. Thus, the repeated comparisons of all subjects were summarized in the analysis below, yielding statistics using 24 answers for each condition and listening position.

Stereo widening: Fig. 3(a) shows the medians and corresponding confidence intervals of perceived source extent in dotted lines for conditions using stereo. For all listening positions an analysis of variance (ANOVA) reveals the dispersion as a significant factor (probability $> 99.9\%$). However, not all neighboring conditions yield significantly different extent for the off-center positions. At the position $(0, -1)\,\mathrm{m}$, the conditions with $70°$ and $90°$ yield no significantly different mean values (13%). Comparing the median values, the conditions with $70°$ and $90°$ at $(0, -0.5)\,\mathrm{m}$, as well as $50°$ and $70°$ are not significantly different. Thus, the lateral distance to central listening position decreases the controllability of stereo widening. While the perceived differences remain between conditions with small dispersion, differences for the greatest dispersion get lost. Note that each perceptual scale in Fig. 3(a) is relative and covers all the 9 conditions (Stereo and Ambisonics) but is specific to one listening position. The diagram does not permit direct comparison of scale values for different listening positions.

Ambisonics: For the Ambisonic widening, dispersion is still a significant factor ($> 99.9\%$). Furthermore the neighboring conditions yield significantly different mean ($> 95\%$) and median values for all listening positions. Other than before, the nearly linear slope between $20°$ and $65°$ gets steeper and not shallower for off-center positions. This is plausible as both dispersion and lateral proximity of the listening position to LS2 increase lateral sound.



(a) Relative perceptual scales for efficient widening at different listening poitions. (b) $\mathrm{IACC}_{\mathrm{E3}}$ for efficient widening.

Figure 3: Perceptual, relative scales at each listening position for efficient stereo and Ambisonic widening and dependency of measured $1 - \mathrm{IACC}_{\mathrm{E3}}$ on dispersion $\hat{\phi}$. Perceptual scales show medians and corresponding 95% confidence intervals and their values are normalized to the maximum extent at each listening position.

While the perceived extent of stereo and Ambisonic widening is similar at the central listening position, the less controllable extent of stereo widening is strongly reduced at the off-center listening positions when compared to Ambisonic widening.

Inter-aural cross correlation coefficients (IACC) were calculated from binaural impulse responses of a B&K HATS 4128C in the experimental setup. The IACC is defined as the maximum of the inter-aural cross correlation function (IACF), cf. [15]. This article uses the $\mathrm{IACC}_{\mathrm{E3}}$, which is the average of the early IACCs for three octave bands around 500 Hz, 1 kHz, and 2 kHz, cf. [16]. For all the 9 conditions, the $\mathrm{IACC}_{\mathrm{E3}}$ values are correlated to the perceived source extent by 91% at $(0, 0)$ m, and by 92% for the off-center listening position at $(0, -0.5)$ m and 82% at $(0, -1)$ m, cf. Tab. 1 and Fig. 3(b). Neither relative scales nor correlations inter-relate the perceived extent at different listening positions.

4. WIDENING AND DIFFUSION FOR MASTERING

An obvious extension of the algorithm is an enlarged time grid $\mathrm{Q}/f_s \geq 10\,\mathrm{ms}$. For dispersion $\hat{\phi} \approx 80°$, this causes a diffuse spatial reverb characterized by an equally long attack and decay. Hereby, relatively short diffuse reverberation can be created.

Below, efficient Ambisonic widening/diffusion is extended to yield high-quality 2D and 3D mastering effects, see Figs. 4 and 6.

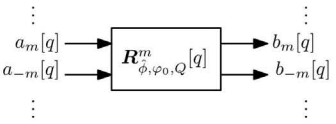
4.1. Frequency-dependent rotation.

The widening algorithm is applicable in mastering of Ambisonic productions or recordings after re-expressing it as a rotation. Using $\cos[\hat{\phi}\cos\Omega + \varphi_0] = \sum_{l=-\infty}^{\infty} \cos(\frac{\pi}{2}\,|l| + \varphi_0)\; J_{|l|}(\hat{\phi})\, e^{\mathrm{i}l\Omega}$ to define frequency-dependent Ambisonic z-rotation matrices achieves the effect. All Ambisonic signals $a_{\pm m}[q]$ belonging to cos and sin harmonics of same m are rotated in pairs by convolution with 2×2 matrices, see Fig. 4(a),

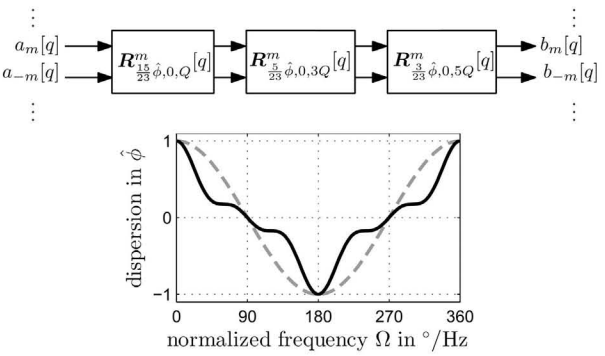
$$\begin{bmatrix} b_m[q] \\ b_{-m}[q] \end{bmatrix} = \underbrace{\begin{bmatrix} \cos(m\,\zeta) & -\sin(m\,\zeta) \\ \sin(m\,\zeta) & \cos(m\,\zeta) \end{bmatrix}}_{:=\boldsymbol{R}^m_{\hat{\phi},\varphi_0,\mathrm{Q}}[q]} \star \begin{bmatrix} a_m[q] \\ a_{-m}[q] \end{bmatrix}. \qquad (11)$$

Setting the rotation angle to $\zeta = \hat{\phi}\,\cos(\Omega) + \varphi_0$ creates a time-domain 2×2 filter matrix containing the sparse responses of Eq. (3)

$$\boldsymbol{R}^m_{\hat{\phi},\varphi_0,\mathrm{Q}}[\lambda\mathrm{Q}] = \begin{bmatrix} h(m\hat{\phi}, m\varphi_0, (\lambda-\Lambda)\mathrm{Q}] & h(m\hat{\phi}, m\varphi_0 + \frac{\pi}{2}, (\lambda-\Lambda)\mathrm{Q}] \\ h(m\hat{\phi}, m\varphi_0 - \frac{\pi}{2}, (\lambda-\Lambda)\mathrm{Q}] & h(m\hat{\phi}, m\varphi_0, (\lambda-\Lambda)\mathrm{Q}] \end{bmatrix}.$$

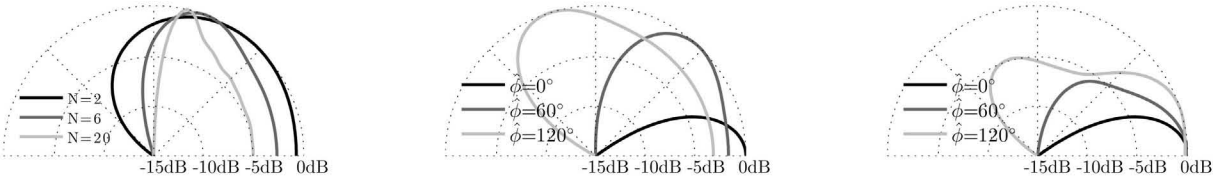


(a) scheme for rotational widening/diffusion.



(b) scheme for depicted, alternative dispersion curve.

Figure 4: (a) processing for dispersive rotation, (b) chained up to produces alternative dispersion curves.



(a) cos-shaped widening at dispersion $\hat{\phi} = 80°$ and varying Ambisonic orders.

(b) cos-shaped widening at varying dispersion and Ambisonic order $\mathrm{N} = 5$.

(c) harmonic-series-shaped widening at varying dispersion and Ambisonic order $\mathrm{N} = 5$.

Figure 5: Broad-band energy distribution $E(\varphi)$ for widening drawn over angle: (a) Ambisonic widening with cosine curve and varying orders, (b) for fifth order and varying dispersion, (c) for fifth order and modified dispersion curve.

4.2. Chaining-up rotations for other dispersion curves

The cos-shaped dispersion around $\varphi_0 = 0$ yields accumulation points in its angularly symmetric broad-band energy distribution

$$E(\varphi) = \int_0^{2\pi} \left\{ \sum_{m=0}^{\mathrm{N}} \cos[\tfrac{m\pi}{2(\mathrm{N}+1)}] \, \cos[m(\varphi - \hat{\phi}\cos\Omega)] \right\}^2 \mathrm{d}\Omega,$$

especially in higher-order Ambisonics, cf. Fig. 5(a). Fig. 4(b) shows a harmonically chained-up rotational dispersion to obtain the curve $\zeta = \sum_{j=0}^{2} \frac{15\,\hat{\phi}}{23(2j+1)} \cos[(2j+1)\Omega)]$,

$$\hat{\boldsymbol{R}}^m_{\hat{\phi},0,\mathrm{Q}}[q] = \boldsymbol{R}^m_{\frac{15}{23}\hat{\phi},0,\mathrm{Q}}[q] \star \boldsymbol{R}^m_{\frac{5}{23}\hat{\phi},0,3\mathrm{Q}}[q] \star \boldsymbol{R}^m_{\frac{3}{23}\hat{\phi},0,5\mathrm{Q}}[q]. \quad (12)$$

The resulting energy distribution in Fig. 5(c) shows less prominent extrema and a stable center compared to Fig. 5(b).

4.3. Application to 3D Ambisonics

The same 2×2 filter matrices are involved when dispersively rotating 3D Ambisonic signals $a_{n,\pm m}$ around the z axis, because the signals are related to the same azimuth harmonics as in 2D. We set $\varphi_0 = 0$ and use

$$\begin{bmatrix} b_{n,m}[q] \\ b_{n,-m}[q] \end{bmatrix} = \boldsymbol{R}^m_{\hat{\phi},0,\mathrm{Q}}[q] \star \begin{bmatrix} a_{n,m}[q] \\ a_{n,-m}[q] \end{bmatrix}. \quad (13)$$

However, sounds close to the z axis remain entirely unaffected.

As a remedy, rotational dispersion can be sequentially applied around z, y, and x. The same filter matrix is involved after using static rotations intermediately aligning y or x with the vertical direction. Given such rotation matrices $\boldsymbol{R}^{(\cdot z)}$ and $\boldsymbol{R}^{(z\cdot)}$, we define:

$$\begin{aligned} \boldsymbol{R}^{(z)}_{\hat{\phi},0,\mathrm{Q}}[q] &= [\boldsymbol{R}^m_{\hat{\phi},0,\mathrm{Q}}[q]]_{n,m}, \quad (14)\\ \boldsymbol{R}^{(y)}_{\hat{\phi},0,\mathrm{Q}}[q] &= \boldsymbol{R}^{(yz)}\, \boldsymbol{R}^{(z)}_{\hat{\phi},0,\mathrm{Q}}[q]\, \boldsymbol{R}^{(zy)}, \\ \boldsymbol{R}^{(x)}_{\hat{\phi},0,\mathrm{Q}}[q] &= \boldsymbol{R}^{(xz)}\, \boldsymbol{R}^{(z)}_{\hat{\phi},0,\mathrm{Q}}[q]\, \boldsymbol{R}^{(zx)}, \end{aligned}$$

and execute to manipulate the set of Ambisonic signals $\boldsymbol{a}[q]$

$$\boldsymbol{b}[q] = \boldsymbol{R}^{(x)}_{\hat{\phi},0,1.9\mathrm{Q}}[q] \star \boldsymbol{R}^{(y)}_{\hat{\phi},0,1.3\mathrm{Q}}[q] \star \boldsymbol{R}^{(z)}_{\hat{\phi},0,\mathrm{Q}}[q] \star \boldsymbol{a}[q]. \quad (15)$$

Different time grids $\{1.9\mathrm{Q}, 1.3\mathrm{Q}, \mathrm{Q}\}$ for x, y, and z achieve a trajectory over frequency that evenly fills a spherical cap of the angular size $\approx 2\hat{\phi}$, as shown in Fig. 6.

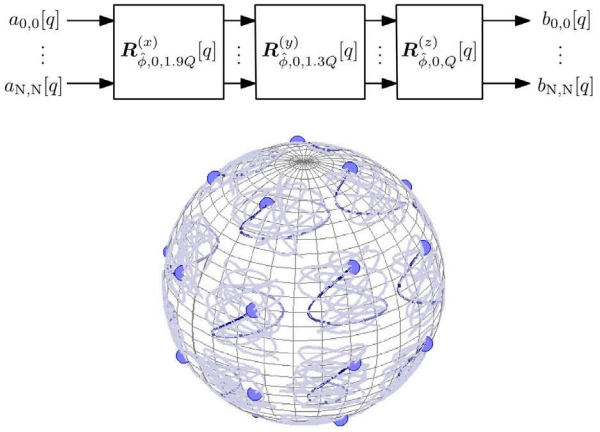


Figure 6: Scheme and trajectories for 3D widened/diffused providing sphere-cap-shaped dispersion.

5. CONCLUSION AND OUTLOOK

A highly effective Ambisonic encoding system for phantom source widening has been presented that disperses the direction of arrival of the sound over frequency.

Listening experiments for horizontal second-order Ambisonics have successfully proven that the new algorithm not only efficiently but also effectively controls the perceived source extent, which can be described by the $\mathrm{IACC}_{\mathrm{E3}}$.

Remarkably good control was also gained outside the central listening spot, where the effect of the new algorithm outperforms its stereophonic counterpart, even for a 1 m displacement.

In higher orders, the simple version of the algorithm might yield perceivable accumulation points of the distributed sound at the turning points of the angular dispersion curve. An alternative was given using a series of rotational dispersions of different frequency normalization / time grid Q.

Moreover, the algorithm can be generalized to work as 3D Ambisonic rotation matrix dispersing around all the three Cartesian axes.

It is subject to future research to investigate simple mathematical models of the decorrelation and widening for multi-channel audio, see also [17]. In particular, investigation of the matrix of inter-channel correlation coefficients appears promising. In addition, it is a perceptually interesting question how the rotational widening algorithm performs (a) for different dispersion contours

and (b) for rotation along other than the vertical axis. To assess how far coloration is a problem, a promising strategy of evaluation is to regard the binaural composite loudness levels, cf. [18].

6. REFERENCES

[1] Manfred R. Schröder, "An artificial stereophonic effect obtained from a single audio signal," *J. Audio Eng. Soc.*, vol. 6, no. 2, 1958.

[2] Benjamin B. Bauer, "Phasor analysis of some stereophonic phenomena," *The Journal of the Acoustical Society of America*, vol. 33, no. 11, pp. 1536–1539, 1961.

[3] Michael A. Gerzon, "Stereophonic signal processor generating pseudo stereo signals," *Patent, WO 93/25055*, 1993.

[4] Gary S. Kendall, "The decorrelation of audio signals and its impact on spatial imagery," *Computer Music J.*, vol. 19, no. 4, 1995.

[5] Maurice Bouéri and Chris Kyriakakis, "Audio signal decorrelation based on a critical band approach," in *Preprint 6291, 117th Conv. Audio Eng. Soc.*, San Francisco, 2004.

[6] Guillaume Potard, *3D-audio object oriented coding*, Ph.D. thesis, University of Wollongong, 2006.

[7] Franz Zotter, Matthias Frank, Georgios Marentakis, and Alois Sontacchi, "Phantom source widening with deterministic frequency dependent time delays," in *DAFx-11*, 2011.

[8] Franz Zotter and Matthias Frank, "Efficient phantom source widening," *Archives of Acoustics*, vol. 38, no. 1, pp. 27–37, 2013.

[9] Mikko-Ville Laitinen, Tapani Philajamäki, Cumhur Erkut, and Ville Pulkki, "Parametric time-frequency representation of spatial sound in virtual worlds," *ACM Trans. Appl. Percept.*, vol. 9, no. 2, 2012.

[10] Jung-Woo Choi, "Source-width extension technique for sound field reproduction systems," in *Proc. 52nd Int. AES Conf.: Sound Field Control - Engineering and Perception, Guildford*, Sept. 2013.

[11] John R. Carson, "Notes on the theory of modulation," *Proceedings of the Institute of Radio Engineers*, vol. 10, no. 1, pp. 57–64, 1922.

[12] ITU, "ITU-R BS.1116-1: Methods for the subjective assessment of small impairments in audio systems including multichannel sound systems," 1997.

[13] Jérôme Daniel, *Représentation de champs acoustiques, application à la transmission et à la reproduction de scénes sonores complexes dans un contexte multimédia*, Phd thesis, Université Paris 6, 2001.

[14] EBU, "Tech 3253 - Sound Quality Assessment Material (SQAM)," http://tech.ebu.ch/publications/sqamcd.

[15] ISO, "ISO 3382-1:2009: Acoustics - measurement of room acoustic parameters - part 1: Performance spaces," 2009.

[16] Takayuki Hidaka, Leo L. Beranek, and Toshiyuki Okano, "Interaural cross-correlation, lateral fraction, and low- and high-frequency sound levels as measures of acoustical quality in concert halls," *The Journal of the Acoustical Society of America*, vol. 98, no. 2, pp. 988–1007, 1995.

[17] Matthias Frank and Franz Zotter, "Simple technical prediction of phantom source widening," in *AIA/DAGA, Fortschritte der Akustik*, Meran, 2013.

[18] Kazuho Ono, Ville Pulkki, and Matti Karjalainen, "Binaural modeling of multiple sound source perception: Methodology and coloration experiments," in *Audio Engineering Society Convention 111*, 11 2001.

[19] "NIST Digital Library of Mathematical Functions," http://dlmf.nist.gov/, Release 1.0.6 of 2013-05-06.

[20] F. W. J. Olver, D. W. Lozier, R. F. Boisvert, and C. W. Clark, Eds., *NIST Handbook of Mathematical Functions*, Cambridge University Press, New York, NY, 2010.

7. APPENDIX

In order to observe the Fourier components of the $\cos[\alpha\cos(\Omega)]$ and $\sin[\alpha\cos(\Omega)]$ functions above, we can regard

$$\begin{aligned}\cos[\alpha\cos(\Omega)+\beta] &= \frac{1}{2}\left\{e^{\mathrm{i}[\alpha\sin(\Omega)+\beta]} + e^{-\mathrm{i}[\alpha\cos(\Omega)+\beta]}\right\} \\ &= \underbrace{\frac{e^{\mathrm{i}\beta}e^{\mathrm{i}\frac{\alpha}{2}[e^{\mathrm{i}\Omega}+e^{-\mathrm{i}\Omega}]}}{2}}_{:=f(\beta,\alpha,\Omega)} + \underbrace{\frac{e^{-\mathrm{i}\beta}e^{-\mathrm{i}\frac{\alpha}{2}[e^{\mathrm{i}\Omega}+e^{-\mathrm{i}\Omega}]}}{2}}_{f^\star(\beta,\alpha,\Omega)=f(-\beta,-\alpha,\Omega)},\end{aligned}$$

which is further, using $J_\lambda(\alpha) = (-1)^\lambda J_{-\lambda}(\alpha)$, [19, 20, Eq. 10.4.1],

$$\begin{aligned}f(\beta,\alpha,\Omega) = \frac{e^{\mathrm{i}\beta}}{2}e^{\mathrm{i}\frac{\alpha}{2}[e^{\mathrm{i}\Omega}+e^{-\mathrm{i}\Omega}]} &= \frac{e^{\mathrm{i}\beta}}{2}\sum_{\lambda=-\infty}^{\infty}\mathrm{i}^\lambda\, J_\lambda(\alpha)\, e^{\mathrm{i}\lambda\Omega} \\ &= \frac{e^{\mathrm{i}\beta}}{2}\sum_{\lambda=-\infty}^{\infty}\mathrm{i}^{|\lambda|}J_{|\lambda|}(\alpha)\, e^{\mathrm{i}\lambda\Omega}.\end{aligned}$$

Using $J_\lambda(-\alpha) = (-1)^\lambda J_\lambda(\alpha)$, [19, 20, Eq. 10.11.1], we obtain

$$\begin{aligned}f(-\beta,-\alpha,\Omega) = = \frac{e^{-\mathrm{i}\beta}}{2}\sum_{\lambda=-\infty}^{\infty}\mathrm{i}^{|\lambda|}J_{|\lambda|}(-\alpha)\, e^{\mathrm{i}\lambda\Omega} \\ = \frac{e^{-\mathrm{i}\beta}}{2}\sum_{\lambda=-\infty}^{\infty}\mathrm{i}^{-|\lambda|}J_{|\lambda|}(\alpha)\, e^{\mathrm{i}\lambda\Omega},\end{aligned}$$

and are finally able to re-expand $\cos[\alpha\cos\Omega+\beta]$ in terms of

$$\begin{aligned}\cos[\alpha\cos\Omega+\beta] &= \frac{1}{2}\sum_{\lambda=-\infty}^{\infty}\left[e^{\mathrm{i}\beta}\,\mathrm{i}^{|\lambda|} + e^{-\mathrm{i}\beta}\,\mathrm{i}^{-|\lambda|}\right] J_{|\lambda|}(\alpha)\, e^{\mathrm{i}\lambda\Omega} \\ &= \sum_{\lambda=-\infty}^{\infty}\cos(\beta+\tfrac{\pi}{2}\,|\lambda|)\; J_{|\lambda|}(\alpha)\, e^{\mathrm{i}\lambda\Omega}. \qquad (16)\end{aligned}$$

Transformation into the time domain just uses the transform pair $e^{\mathrm{i}\lambda\Omega} \leftrightarrow \delta(t-\lambda)$ after frequency scaling $e^{\mathrm{i}\lambda\mathrm{T}\Omega} \leftrightarrow \delta(t-\lambda\mathrm{T})$.

MEASUREMENT-BASED MODAL BEAMFORMING USING PLANAR CIRCULAR MICROPHONE ARRAYS

Markus Zaunschirm
Institute of Electronic Music and Acoustics
Univ. of Music and Performing Arts Graz
Graz, Austria
zaunschirm@iem.at

Franz Zotter
Institute of Electronic Music and Acoustics
Univ. of Music and Performing Arts Graz
Graz, Austria
zotter@iem.at

ABSTRACT

This paper describes how to use a planar circular pressure-zone table-top microphone array for modal beamforming. Its goals are similar as for spherical arrays: higher-order resolution and a more-or-less steering-invariant beampattern design in the three-dimensional half space. As conventional circular arrays lack control of the beampattern in the vertical array plane, the proposed arrangement tries to fix this shortcoming to allow both horizontal and vertical control of beamforming. To provide a fully calibrated decomposition into the directional modes, the proposed beamforming approach is based on measurement data. From a MIMO (multiple-input-multiple-output) system description of the measurement data in the spherical harmonics domain, an inverse MIMO system of filters is designed for decomposing the microphone array signals into those spherical components eligible for modal beamforming. For an efficient measurement and robust set of decomposition filters, a reduced set of measurement positions and a regularisation strategy is suggested.

1. INTRODUCTION

Beamforming denotes the discrimination between signals based on the spatial location of sources. Whilst conventional beamforming algorithms directly operate on the sensor signals, modal beamforming approaches use directional modes that are obtained by decomposing the wavefield into orthogonal solutions of the acoustic wave equation. Overviews are given in [1, 2].

Spherical arrays are most generic, but also require a lot of hardware effort. In real-world scenarios where acoustic sources are restrained to the upper half of the three-dimensional space the geometry of the microphone array needs to be adopted. For example Li and Duraiswami [3] designed a hemispherical table microphone for sound capture and beamforming. In order to further decrease cost and hardware complexity, circular planar microphone arrays are feasible. In [4] a planar circular table microphone array, consisting of three near-coincident cardioid microphones, is presented that allows for decomposition of the acoustic scene in modes of first order. Using this setup looses its directivity when steering the spatial sensitivity into the vertical direction. In order to improve the spatial selectivity of the generated beampattern, a higher order resolution is required. An approach for a decomposition of the soundfield in second-order directional modes is outlined in literature by Meyer and Elko in [5] and [6]. They suggest a microphone array consisting of omnidirectional microphones on a concentric circle and an omnidirectional centre element. Craven et al [7] argue that acoustic gradient sensors are preferable in signal to noise behaviour as they decrease the bass boosts of array processing filters. On the other hand manufacturing of cardioid microphones usually does not yield capsules that are as well-matched as omnidirectional ones, although a precise match is required for accurate array processing. Additionally, the orientation of the cardioid microphones needs to be precise in order to avoid mismatch between analytic model and prototype. This paper describes how to obtain a fully calibrated decomposition into the directional modes, and a beamforming approach that is entirely based on measurement data and does not rely on an analytic model. A schematic layout of the used microphone array is depicted in figure 1. The central element is an omnidirectional microphone and five equi-spaced outwards-oriented cardioid microphones lie on a concentric circle of the radius r. The array is planar and intended for use as a pressure-zone microphone that is placed for example on table. Possible applications include beamforming for teleconferencing and to a certain extent also 3D spatial recordings.

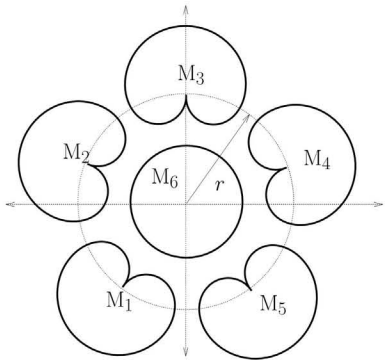


Figure 1: *Schematic layout of the prototype.*

2. MEASUREMENTS OF MICROPHONE ARRAY CHARACTERISTICS

The directional sensitivity of the six array microphones can be measured for one direction by recording a sweep from a loudspeaker that is placed there. For a complete measurement of as many directional responses as possible, the direction-dependent response of each microphone is measured by a hemispherically surrounding loudspeaker array.

Measurement excitation positions (loudspeaker positions) are set according to a spatial resolution of $\Delta\varphi = 10^\circ$ in azimuth and $\Delta\vartheta = 11.25^\circ$ in zenith direction which leads to a grid layout of 8

latitude circles and 36 meridians. The right half of fig. 2 depicts the hemispherical measurement configuration of 288 loudspeakers on a radius of $1.3m$. In order to reduce the measurement effort, the number of directions may be reduced to 6 instead of 288, see red dots in the right half of fig. 2. In the practical setup, fig. 2, an eight-element quarter-circular loudspeaker array could be used for sequentially measuring 8 directions of different zenith angles at one azimuth angle. In order to reach all 36 proposed azimuth angles, the microphone array is rotated by a computer-controlled turntable between the measurements.

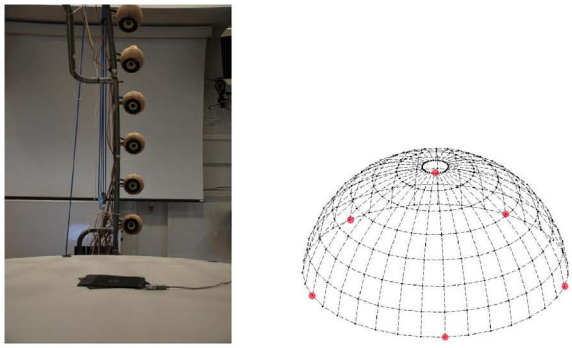


Figure 2: *Setup for capturing responses on a* 8×36 *longitude/latitude grid; red dots mark the on-axis directions.*

The measurements span an 8×36 set of impulse responses $h_{\lambda\mu}[\tau]$, where λ, μ and τ are the loudspeaker, microphone and discrete-time indices, respectively.

The actual measurements are done according to the exponentially swept-sine (ESS) method presented by Farina in [8]. The impulse responses between the λ^{th} loudspeaker to the μ^{th} microphone are calculated in the frequency domain by a simple division of the two spectra defined in eq. (1)

$$h_{\lambda\mu}[\tau] = IFFT\left[\frac{FFT(x_{\lambda\mu}[\tau])}{FFT(s[\tau])}\right], \tag{1}$$

where $x_{\lambda\mu}[\tau]$ and $s[\tau]$ denote the recorded response and the exponential sweep, respectively. The influences of the loudspeaker characteristics are minimized by equalizing according to a calibrated reference microphone placed in the centre of the experimental setup.

2.1. Directivity patterns

In this section, we analyse the measurement data in order to make a statement about the three dimensional directivity patterns and their rotational symmetry of the analysed microphones. Let us define a vector that contains the discrete directional response of the μ^{th} microphone ($\mu = 1, \ldots, 6$) as

$$\boldsymbol{h}_\mu(\omega) = \begin{pmatrix} h_\mu(\boldsymbol{\theta}_1, \omega) \\ h_\mu(\boldsymbol{\theta}_2, \omega) \\ \vdots \\ h_\mu(\boldsymbol{\theta}_\mathrm{L}, \omega) \end{pmatrix}, \tag{2}$$

where $\boldsymbol{\theta}_\lambda = [\cos(\varphi_\lambda)\sin(\vartheta_\lambda),\ \sin(\varphi_\lambda)\sin(\vartheta_\lambda),\ \cos(\vartheta_\lambda)]^\mathrm{T}$ is a direction vector using the azimuth and zenith angle φ_λ and ϑ_λ of the λ^{th} loudspeaker ($\lambda = 1, \ldots, \mathrm{L}$) in spherical coordinates. The magnitude of the directivity pattern at a specific frequency is then plotted using a spherical meshgrid of the available grid positions in fig. 3.

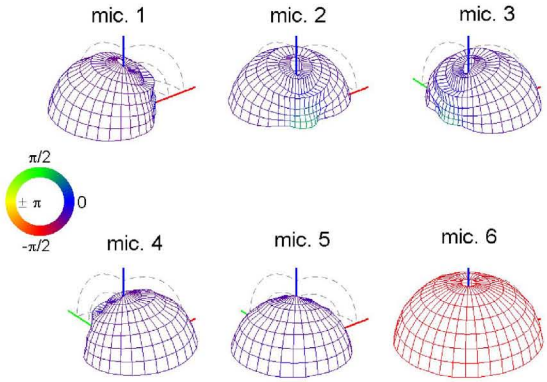


Figure 3: *Directivity patterns of mounted microphones at* $f \approx 1\,\mathrm{kHz}$ *on the* 8×36 *measurement grid; color expresses the phase.*

2.2. Directivity patterns in modal domain

From the series of L loudspeaker responses to the μ^{th} microphone, we obtain a single response by linear combination of all loudspeakers with the weight $\boldsymbol{g} = [g_1, \ldots, g_\mathrm{L}]^\mathrm{T}$

$$h_\mu(\omega) = \boldsymbol{h}_\mu^\mathrm{T}(\omega)\,\boldsymbol{g}. \tag{3}$$

As a counterpart to the weight vector of the spaced loudspeakers at their discrete locations, we define a continuous driving distribution $g(\boldsymbol{\theta})$ depending on the direction vector $\boldsymbol{\theta}$. Such a function is related to the loudspeaker weights g_λ by

$$g(\boldsymbol{\theta}) = \sum_{\lambda=1}^{\mathrm{L}} \delta(\boldsymbol{\theta} - \boldsymbol{\theta}_\lambda)\, g_\lambda, \tag{4}$$

$\delta(\boldsymbol{\theta} - \boldsymbol{\theta}_\lambda)$ symbolizing the directional Dirac delta function. For a modal representation, the equation is expanded in spherical harmonics $Y_n^m(\boldsymbol{\theta})$

$$g(\boldsymbol{\theta}) = \sum_{n=0}^{\infty} \sum_{m=-n}^{n} \underbrace{\sum_{\lambda=1}^{\mathrm{L}} Y_n^m(\boldsymbol{\theta}_\lambda)\, g_\lambda}_{:=\gamma_n^m}\, Y_n^m(\boldsymbol{\theta}), \tag{5}$$

$Y_n^m(\boldsymbol{\theta}_\lambda)$ being the expansion coefficients of the Dirac delta function and γ_{nm} of their weighted superposition. The resolution of $g(\boldsymbol{\theta})$ is limited by truncating the summation in n to $n \leq \mathrm{N}$, a number of functions we can resolve by the measurement loudspeakers. In terms of matrices and vectors, we may write instead of $\gamma_n^m = \sum_{\lambda=1}^{\mathrm{L}} Y_n^m(\boldsymbol{\theta}_\lambda)\, g_\lambda$

$$\begin{aligned} \boldsymbol{\gamma}_\mathrm{N} = \boldsymbol{L}_\mathrm{N}\,\boldsymbol{g}, \qquad & \text{with } \boldsymbol{L}_\mathrm{N} = [\boldsymbol{y}_\mathrm{N}(\boldsymbol{\theta}_1), \ldots, \boldsymbol{y}_\mathrm{N}(\boldsymbol{\theta}_\mathrm{L})], \\ & \boldsymbol{y}_\mathrm{N}^\mathrm{T}(\boldsymbol{\theta}) = [Y_0^0(\boldsymbol{\theta}), \ldots, Y_\mathrm{N}^\mathrm{N}(\boldsymbol{\theta})], \\ & \text{and } \boldsymbol{\gamma}_\mathrm{N}^\mathrm{T} = [\gamma_0^0, \ldots, \gamma_\mathrm{N}^\mathrm{N}]. \end{aligned} \tag{6}$$

If we wish to create specific SH coefficients $\boldsymbol{\gamma}_\mathrm{N}$ with the loudspeakers, their weights should be $\boldsymbol{g} = \boldsymbol{L}_\mathrm{N}^+\,\boldsymbol{\gamma}_\mathrm{N}$, using the pseudo-inverse

of $\boldsymbol{L}_\mathrm{N}$. Choosing $\boldsymbol{\gamma}_\mathrm{N} = \boldsymbol{y}_\mathrm{N}(\boldsymbol{\theta})$, we obtain the sensitivity of the μ^{th} microphone interpolated in terms of SH:

$$h_\mu(\boldsymbol{\theta}, \omega) = \boldsymbol{h}_\mu^\mathrm{T}(\omega)\, \boldsymbol{L}_\mathrm{N}^+\, \boldsymbol{y}_\mathrm{N}(\boldsymbol{\theta}). \tag{7}$$

Note that here some spherical harmonics need to be excluded from the vector $\boldsymbol{y}_\mathrm{N}(\boldsymbol{\theta})$. Because the microphone array under test is a table-top pressure-zone array, we are given a sound-rigid acoustic boundary condition on the horizontal plane. Only the $\frac{(\mathrm{N}+1)(\mathrm{N}+2)}{2}$ even-symmetric spherical harmonics $Y_n^{2s-n}(\boldsymbol{\theta})$, $0 \le s \le n$, fulfill this condition; others are grayed out in fig.4 showing $0 \le n \le 2$. For fine interpolation, the given grid allows to choose $\mathrm{N} = 14$ at a

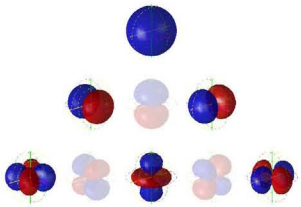


Figure 4: *SHs up to order* 2 *that satisfy the boundary condition (colored); skew symmetric are transparent.*

reasonable condition number for pseudo inversion. Figure 5 shows the three-dimensional directivity pattern of microphone 3 at about 1kHz on a fine grid of 12000 nodes.

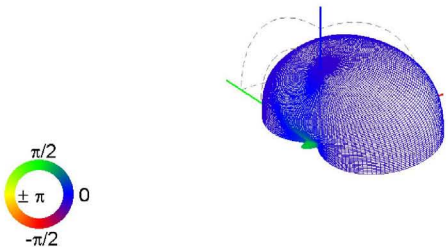


Figure 5: *SH-interpolated directivity pattern of microphone* 3 *at* $f \approx$ 1kHz*; color expresses the phase.*

3. MODAL BEAMFORMING BY MIMO INVERSION OF MEASURED SH DIRECTIVITIES

The targeted spherical harmonic modal beamformer is shown in fig. 6. It processes the microphone signals as to produce a set of spherical harmonic pickup patterns of limited order n, $n \le 2$ in our particular case. This step is somewhat elegant as an output directivity can be formed by subsequent frequency-independent linear combination thereof [1].

3.1. Decomposer Unit

The goal is to design a unit that transforms the recorded microphone signals into the spherical harmonic spectrum γ_n^m, which are the target signals for spherical-harmonic-based modal beamforming.

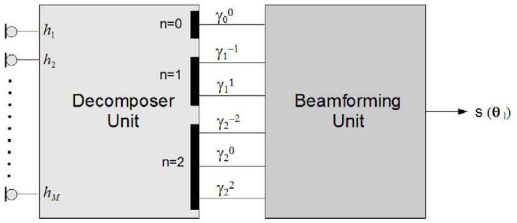


Figure 6: *Scheme of modal beamforming stages;* $s(\boldsymbol{\theta}_\mathrm{l})$ *denotes the output signal for a beam steered towards* $\boldsymbol{\theta}_\mathrm{l}$.

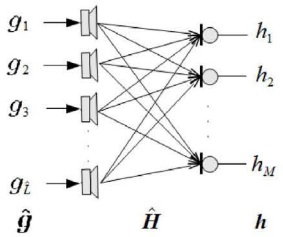


Figure 7: *MIMO system.*

The multiple-input-multiple-output system (MIMO, see fig. 7) of the device under test is described as

$$\boldsymbol{h}(\omega) = \hat{\boldsymbol{H}}(\omega)\, \hat{\boldsymbol{g}}, \tag{8}$$

where $\hat{\boldsymbol{H}}(\omega)$ represent measured MIMO responses, and $\boldsymbol{h}(\omega)$ are now all the M microphone responses due to the loudspeaker weights $\hat{\boldsymbol{g}}$. Omitting (ω) for brevity, the matrix $\hat{\boldsymbol{H}}$ contains responses from Eq. (2)

$$\hat{\boldsymbol{H}} = [\hat{\boldsymbol{h}}_1, \ldots, \hat{\boldsymbol{h}}_\mathrm{M}]^\mathrm{T}. \tag{9}$$

Hatted variables $\hat{\boldsymbol{h}}_\mu(\omega)$ are used to denote a coarse selection of measurements out of $\boldsymbol{h}_\mu(\omega)$, see red dots in fig. 2. This is preferable in practical and repeated calibration measurement, so that modal beamformer design only uses the 6 loudspeaker positions $\boldsymbol{\theta}_\mu = [\cos(\varphi_\mu)\sin(\vartheta_\mu),\ \sin(\varphi_\mu)\sin(\vartheta_\mu),\ \cos(\vartheta_\mu)]^\mathrm{T}$ aligned with all 6 pointing directions of the microphone array elements. The fine 288 measurement grid is only used for later verification. We control the reduced loudspeaker weights $\hat{\boldsymbol{g}}$ by the smaller 6×6 matrix $\boldsymbol{Y}^{-1}$ instead of $\boldsymbol{L}^+$,

$$\hat{\boldsymbol{g}} = \boldsymbol{Y}^{-1}\, \boldsymbol{\gamma}_2, \qquad \text{with } \boldsymbol{Y} = [\boldsymbol{y}_2(\boldsymbol{\theta}_1), \ldots, \boldsymbol{y}_2(\boldsymbol{\theta}_\mathrm{M})], \quad \text{and } \boldsymbol{\gamma}_2 = [\gamma_0^0, \ldots, \gamma_2^2]. \tag{10}$$

Insertion in Eq. (8) transforms the loudspeaker side of the MIMO system to a modal representation of limited order $n \le 2$ resolvable by the microphones

$$\boldsymbol{h} = \hat{\boldsymbol{H}}\, \boldsymbol{Y}^{-1} \boldsymbol{\gamma}_2. \tag{11}$$

We may further transform the MIMO system $\hat{\boldsymbol{H}}$ from the microphone side into the spherical harmonics domain by using the same expression $\boldsymbol{Y}^{-1}$, yielding the modal signal outputs $\boldsymbol{\chi}_2$ of the microphone array

$$\boldsymbol{\chi}_2 = \boldsymbol{Y}^{-1}\, \hat{\boldsymbol{H}}\, \boldsymbol{Y}^{-1} \boldsymbol{\gamma}_2. \tag{12}$$

In the underlying analytical model [9], this operation would render the system

$$\boldsymbol{Y}^{-1}\, \hat{\boldsymbol{H}}\, \boldsymbol{Y}^{-1} := \boldsymbol{C} \tag{13}$$

perfectly diagonal. We therefore expect some of the paths in the 6×6 MIMO system $\boldsymbol{C}$ to be vanishing. Indeed, no perfect but a diagonalizing effect on the measured MIMO system is observed in $\boldsymbol{C}$, fig. 8. In order to obtain a correct mapping of the spherical harmonics modes at the loudspeaker side to the modes at the microphone side, a decomposition matrix that is inverse to the transformed MIMO system $\boldsymbol{C}$ is introduced

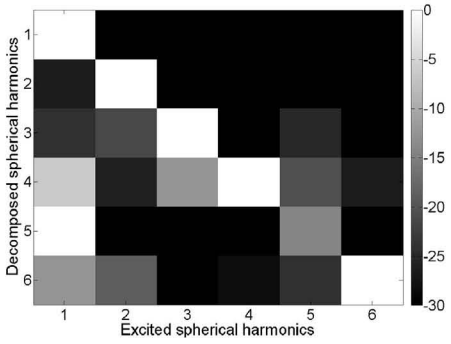


Figure 8: Magnitudes of the MIMO system $\boldsymbol{C}$ for $f = 2\,\text{kHz}$. Values are normalized to a range of 30 dB.

$$\boldsymbol{\chi}_2 = \boldsymbol{D}\,\boldsymbol{C}\,\boldsymbol{\gamma}_2 = \boldsymbol{\gamma}_2. \tag{14}$$

A decomposition matrix $\boldsymbol{D} = \boldsymbol{C}^{-1}$ would yield a perfect but non-robust decomposition of the microphone signals into the mode strengths $\boldsymbol{\chi}_2 = \boldsymbol{\gamma}_2$.

3.2. Regularised inversion of the transformed MIMO system

The transformed MIMO system is square and may be exactly inverted to get $\boldsymbol{D}$ if it is non-singular. The condition number $\kappa(\cdot)$ of a matrix indicates the distance to a regular matrix and is defined as the ratio between the maximal and minimal singular value [10]. A perfectly regular matrix has a condition number $\kappa(\cdot) = 1$, but we expect $\kappa(\boldsymbol{C}) > 1$ in our case. By applying the singular value decomposition (SVD, [11]) on $\boldsymbol{C}$, we obtain

$$\boldsymbol{C} = \boldsymbol{U}\,\boldsymbol{S}\,\boldsymbol{V}^{\mathrm{H}}, \tag{15}$$

where $\boldsymbol{U}$ and $\boldsymbol{V}$ are unitary matrices column-wisely containing the left and right singular vectors of $\boldsymbol{C}$, and $\boldsymbol{S}$ is the diagonal matrix containing the singular values in descending order

$$\boldsymbol{S} = \mathrm{diag}(\boldsymbol{\sigma}), \quad \boldsymbol{\sigma} = [\sigma_1, \ldots, \sigma_N]^{\mathrm{T}}. \tag{16}$$

We define a regularised inverse of the MIMO system matrix $\boldsymbol{C}$ as

$$\boldsymbol{D} = \boldsymbol{V}\tilde{\boldsymbol{S}}^{-1}\boldsymbol{U}^{\mathrm{H}}, \text{with } \tilde{\boldsymbol{S}}^{-1} = \mathrm{diag}(\tilde{\boldsymbol{\sigma}})^{-1}, \tag{17}$$

$$\text{and } \tilde{\boldsymbol{\sigma}} = \boldsymbol{\sigma} + \underbrace{\sigma_1\, c_1}_{\text{local}} + \underbrace{\sigma_{max}\, c_2}_{\text{global}}, \tag{18}$$

where $\tilde{\boldsymbol{\sigma}}$ denotes the regularised singular values, σ_1 denotes the highest singular value of the system matrix at frequency f (local), $\sigma_{max} = \max_f(\sigma_1(f))$ refers to the maximal singular value over the entire frequency range (global), and c_1, c_2 are scalar regularisation constants. They are used to control the amount of regularisation. Local and global regularisation improve the system conditioning to obtain a robust modal decomposer $\boldsymbol{D}$. The local regularisation is used to avoid an extreme amplification of the inverse system at frequencies where just a few components are under-represented in $\boldsymbol{C}$, whereas the global regularisation avoids amplification at frequencies where transfer function components in $\boldsymbol{C}$ are too weak, altogether. Global regularisation is essential in the lower frequency range, whereas local and global regularisation affect the approximation at higher frequencies to the same extent. Fig. 9 shows that regularisation mainly affects the small singular values to improve the robustness of $\boldsymbol{D}$ while accepting it being a less accurate inverse $\boldsymbol{D}\,\boldsymbol{C} \approx \boldsymbol{I}$.

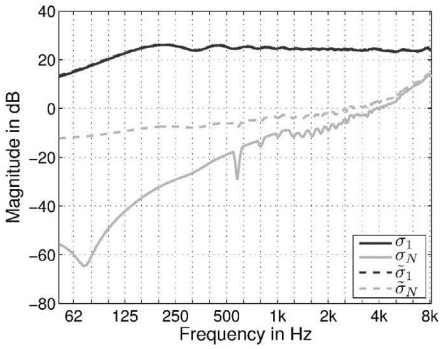


Figure 9: , *Maximal and minimal singular values of the transformed MIMO system and the regularised system for* $c_2 = 0.008$ *and* $c_1 = 0.01$.

3.3. Modal Beamforming Unit

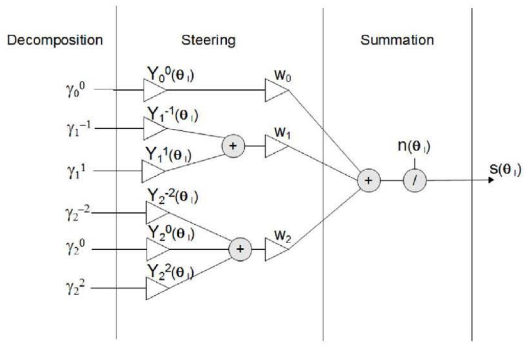


Figure 10: *Scheme of beamforming unit.*

A block diagram of the beamformer unit is depicted in fig.10. Thereby, the input signals are the unified modes in the spherical harmonics domain γ_n^m that are generated in the decomposer unit using the MIMO filter $\boldsymbol{D}$. In the steering unit, these signals are weighted with the spherical harmonics evaluated at the lookdirection $\boldsymbol{\theta}_l$, and in a next step they are multiplied with frequency independent order weights w_n, e.g. the max-r_{E} weights $[1, 0.775, 0.4]$, that are designed to form specific beampattern shapes (see [12], [13], [14] and [9]). In the last unit, the summation unit, the obtained signals are summed up. The beampattern is normalized by its lookdirection amplitude $n(\boldsymbol{\theta}_l) = \sum_{n=0}^{2}\sum_{s=0}^{n} w_n\,[Y_n^{2s-n}(\boldsymbol{\theta}_l)]^2$ to remove its dependency on the zenith angle.

4. RESULTING BEAMPATTERNS

The horizontal slices of the achievable beampatterns with max-r_E order weighting $w_n = [1, 0.775, 0.4]$ steered towards the horizontal array plane are shown in fig. 11. It is striking that the

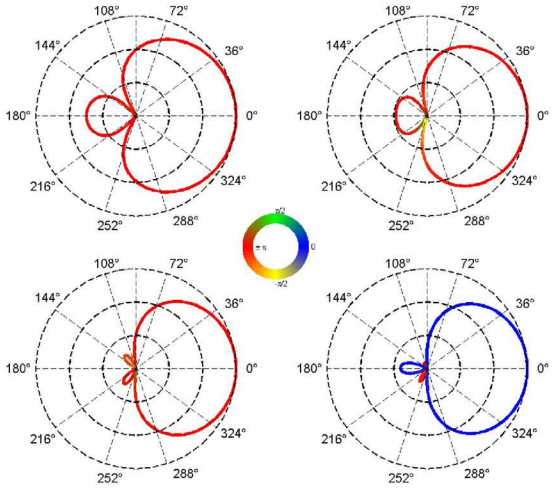


Figure 11: *Horizontal slice of supercardioid beampattern steered toward $z = 0$ at frequencies of $[400, 1000, 2000, 4000]$ Hz and usage of regularised decomposition filters; radial divisions are 10 dB steps.*

obtained directional characteristics are frequency dependent and that the beampatterns evolve from a nearly first order supercardioid at low frequencies to a second order supercardioid characteristic at high frequencies. The vertical slices of a hypercardioid beam ($w_n = [1, 1, 1]$) steered towards the z-direction are shown in fig. 12 where one can observe a similar behaviour as for the horizontal slices, namely that the higher order pattern is just available at higher frequencies. The quasi order-limited beamforming is caused by the

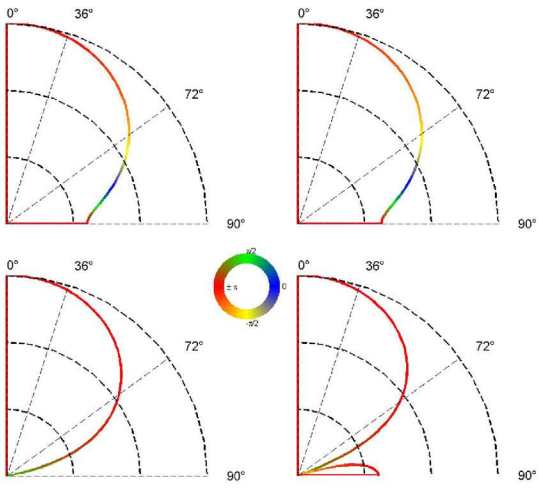


Figure 12: *Vertical slice of hypercardioid beampattern steered towards the z-direction at frequencies of $[400, 1000, 2000, 4000]$ Hz and usage of regularised filters; radial divisions are 10 dB steps.*

filter regularisation, as the higher-order modes are weakly present at low frequencies. This order limitation improves the white noise gain (WNG) and accordingly the robustness of the beamforming system [15].

4.1. Does every array need calibration?

The filters designed by limiting the condition number to ~ 30dB (see fig. 9) still yield a robust decomposition if the array parameters deviate within 1%. But is the MIMO decomposer also applicable to different copies of the microphone array?

In order to test the portability of the regularised decomposition filters to an array duplicate with same geometry, we generated filters with the circularly rotated data of the measured matrix $\hat{\boldsymbol{H}}$. The beamforming system for the rotated filters was tested on the original microphone array. The obtained beampatterns of this setup highly vary from the beampatterns produced with the original data, see fig. 13. Thus, the designed filters are not necessarily applicable to arrays where the array characteristics differ from the array prototype. In [9] it is shown that minor mounting errors (about

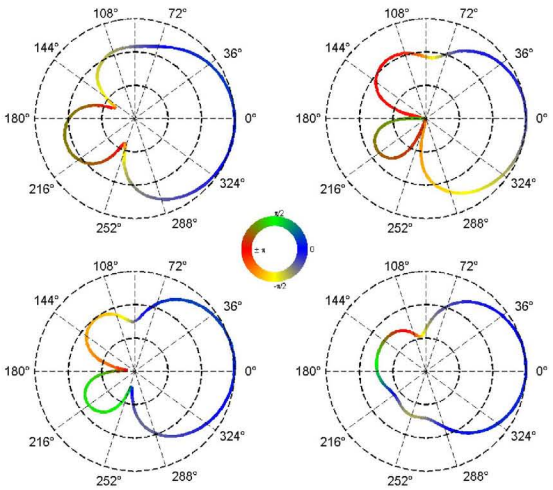


Figure 13: *Horizontal slice of supercardioid beampattern steered toward $z = 0$ at frequencies of $[400, 1000, 2000, 4000]Hz$ and usage of rotated regularised decomposition filters; radial divisions are 10 dB steps.*

$4°$) of the microphone orientation, gain mismatches between the channels as well as deviations of microphone characteristics lead to major decomposition errors. This highlights the importance of a calibration procedure for every microphone array duplicate.

5. CONCLUSIONS

In this paper, we presented and tested a novel modal beamforming approach for a six channel pressure-zone table-top microphone array. The design using five cardioid microphones and one omnidirectional central microphone yields robust frequency responses for creating 2nd order modal beampatterns. A measurement-based decomposition for modal beamforming that exploits the benefits of gradient transducers of reasonable manufacturing accuracy is proposed. The practical approach includes a regularised inversion of a MIMO system and is easy to use as only 6×6 MIMO response

measurements are necessary. The low complexity of this calibration procedure is paramount as exchangeability of the MIMO decomposition filters to array duplicates is not possible for superdirectional beamforming.

6. ACKNOWLEDGMENTS

We want to thank Dominik Biba, Martin Opitz, Richard Pribyl, and Marco Riemann from AKG Acoustics GmbH for the excellent collaboration and for building the microphone array prototype within the project AAP, which was funded by the Austrian ministries BMVIT, BMWFJ, the Styrian Business Agency (SFG), and the departments 3 and 14 of the Styrian Government. The Austrian Research Promotion Agency (FFG) conducted the funding under the Competence Centers for Excellent Technologies (COMET, K-Project), a program of the above mentioned institutions. Further we want to thank Hannes Pomberger (IEM) for co-developing design and algorithm of the presented pressure-zone table-top microphone array.

7. REFERENCES

[1] J. Meyer and G. Elko, "A highly scalable spherical microphone array based on an orthonormal decomposition of the soundfield," in *Acoustics, Speech, and Signal Processing (ICASSP), 2002 IEEE International Conference on*. IEEE, 2002, vol. 2, pp. II–1781.

[2] Heinz Teutsch, *Modal array signal processing: principles and applications of acoustic wavefield decomposition*, vol. 348, Springer, 2007.

[3] Z. Li and R. Ruraiswami, "Hemispherical microphone arrays for sound capture and beamforming," in *Applications of Signal Processing to Audio and Acoustics, 2005. IEEE Workshop on*. IEEE, 2005, pp. 106–109.

[4] F. Reining, "Microphone arrangement comprising pressure gradient transducers," Feb. 23 2009, US Patent App. 12/390,990.

[5] J. Meyer and G. Elko, "Spherical harmonic modal beamforming for an augmented circular microphone array," in *Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on*. IEEE, 2008, pp. 5280–5283.

[6] J.M. Meyer and G.W. Elko, "Augmented elliptical microphone array," July 9 2008, US Patent App. 12/595,082.

[7] Peter G. Craven, Malcolm Law, and Chris Travis, "Microphone array," 2008, Patent App. WO 2008/040991 A2.

[8] A. Farina, "Simultaneous measurement of impulse response and distortion with a swept-sine technique," *Preprints-Audio Engineering Society*, 2000.

[9] Markus Zaunschirm, "Modal beamforming using planar circular microphone arrays," M.S. thesis, University of Music and Performing Arts Graz, 2012.

[10] E.W. Cheney and D.R. Kincaid, *Numerical mathematics and computing*, Brooks/Cole Pub Co, 2007.

[11] J.E. Jackson and J. Wiley, *A user's guide to principal components*, Wiley Online Library, 1991.

[12] J. Daniel, J.B. Rault, and J.D. Polack, "Ambisonics encoding of other audio formats for multiple listening conditions," *PREPRINTS-AUDIO ENGINEERING SOCIETY*, 1998.

[13] Gary W. Elko, "Differential microphone arrays," in *Audio signal processing for next-generation multimedia communication systems*, Yiteng Arden Huang and Jacob Benesty, Eds. Springer, 2004.

[14] Gary W. Elko, "Superdirectional microphone arrays," in *Acoustic signal processing for telecommunication*, Steven L Gay and Jacob Benesty, Eds. Kluwer Academic, 2000.

[15] G.W. Elko, R.A. Kubli, J.M. Meyer, et al., "Audio system based on at least second-order eigenbeams," Sept. 8 2009, US Patent 7,587,054.

Proc. of the EAA Joint Symposium on Auralization and Ambisonics, Berlin, Germany, 3-5 April 2014

RENDERING BINAURAL ROOM IMPULSE RESPONSES FROM SPHERICAL MICROPHONE ARRAY RECORDINGS USING TIMBRE CORRECTION

Jonathan Sheaffer
sheaffer@ee.bgu.ac.il

Shahar Villeval
shaharv@ee.bgu.ac.il

Boaz Rafaely
br@ee.bgu.ac.il

Acoustics Lab,
Dept. of Electrical and Computer Engineering
Ben-Gurion University of the Negev, Israel.

ABSTRACT

The technique of rendering binaural room impulse responses from spatial data captured by spherical microphone arrays has been recently proposed and investigated perceptually. The finite spatial resolution enforced by the microphone configuration restricts the available frequency bandwidth and, accordingly, modifies the perceived timbre of the played-back material. This paper presents a feasibility study investigating the use of filters to correct such spectral artifacts. Listening tests are employed to gain a better understanding of how equalization affects externalization, source focus and timbre. Preliminary results suggest that timbre correction filters improve both timbral and spatial perception.

1. INTRODUCTION

Binaural technology [1] provides a means to render headphone-presented stimuli that mimic sounds as if they were heard in a regular, headphone-free listening situation. It has applications in psychoacoustics research [2], auditory neuroscience [3], architectural acoustics [4] and audio technology [5]. In its simplest form, binaural technology utilizes free field head-related transfer-functions (HRTFs) of a mannikin or an individual subject to spatially filter an audio signal. While such processing provides a basic means to simulate sound localization, the absense of reverberation and time-varying auditory cues have a negative effect with regard to achieving an accurate sense of sound externalization [6, 7].

To obtain a more complete set of auditory cues, one may opt to directly measure the transfer function of a room using a manikin, which results in a Binaural Room Impulse Response (BRIR), e.g. see [8]. This, however, results in a single transfer function combining the effects of the room itself as well as the head, ears and torso, and thus represents the anthropometric features of a specific listener. Additionally, if one wishes to reproduce the effects of head movements, the BRIR needs to be measured in a range of head orientations hence making the measurement procedure inefficient for many practical applications.

Rafaely and Avni [9] suggested a method to render BRIRs in the spherical harmonics (SH) domain, by making use of pre-measured HRTFs and a Spatial Room Impulse Response (SRIR) which can be obtained either by means of numerical simulation or by direct room measurement using a spherical microphone array. More recently, Avni et al. [10] studied the perceptual effects of recording and reproducing sound fields at different spatial resolutions. Among their findings, they discovered that limiting the SH order of the recorded sound field has a prominent effect on the frequency bandwidth of the resulting BRIR, and hence on the timbre of the played-back material. In other words, a BRIR of low spatial resolution also has a limited frequency bandwidth, which indicates that the spatial and spectral design parameters of BRIRs should not be seen in isolation. To address this, Villeval [11] suggested a timbre equalization filter, compensating for the average change in frequency response between BRIRs constructed at two different spatial resolutions.

The results presented in [10] showed that there is an inherent trade-off between the spatial resolution of the sound field recorded with a spherical array and the spectral representation of the resulting BRIR. As a first step towards addressing this trade-off, this paper presents a feasibility study on the effects of correcting low-order BRIRs with a timbre equalization filter. The paper is structured as follows: Sec. 2 and 3 briefly outline the procedure for computing a BRIR in the SH domain, and for equalizing it to a desired SH order. This is followed by experimental results from a preliminary listening test in Sec. 4, which are further discussed in Sec. 5.

2. RENDERING BINAURAL ROOM IMPULSE RESPONSES

To render a BRIR from sound pressure measured by a spherical microphone array, the method suggested in [9] is followed in this paper. Let $H^l(k,\Omega)$ and $H^r(k,\Omega)$ denote a set of pre-measured HRTFs for the left and the right ear, respectively, where $k = 2\pi f/c$ is the acoustic wavenumber, f is the frequency and c is the speed of sound in air. Here, $\Omega \equiv (\theta,\phi) \in S^2$ denotes the angle in a standard spherical coordinate system [12] in which (r,θ,ϕ) denote radial distance, elevation and azimuth, respectively. By applying the spherical Fourier transform [13] to $H^l(k,\Omega)$ and $H^r(k,\Omega)$, one obtains their respective representations in the SH domain, $H^l_{nm}(k)$ and $H^r_{nm}(k)$.

Similarly, let $p(k,r,\Omega)$ denote some pressure function on a sphere that is square integrable over Ω and whose spherical Fourier

transform yields the function $p_{nm}(k,r)$. In a room, this function represents spatial information on a continuum of plane waves arriving at the receiving position from the sound source and the different reflecting surfaces. The complex amplitudes of the spherical harmonic components of these plane waves, $a_{nm}(k)$, can be obtained by performing a plane-wave decomposition of the sound field as follows [14]:

$$a_{nm}(k) = \frac{p_{nm}(k,r)}{b_n(kr)}. \tag{1}$$

In this paper all spherical array measurements are performed directly over the surface of a rigid sphere and, as such, $b_n(kr)$ is given by [15]

$$b_n(kr) = 4\pi i^n \left[j_n(kr) - \frac{j_n'(kr)}{h_n'(kr)} h_n(kr) \right], \tag{2}$$

where $j_n(\cdot)$ is the spherical Bessel function, $h_n(\cdot)$ is the spherical Hankel function and $j_n'(\cdot)$ and $h_n'(\cdot)$ represent their first derivatives with respect to the argument. For high values of n, the result of $b_n(kr)$ approaches zero for low values of kr, which requires a large calculation dynamic range. To overcome this numerical limitation, in this paper $b_n(kr)$ is soft-limited according to the procedure suggested in [16].

Once a plane wave decomposition is performed, a BRIR can be calculated as follows [9]:

$$p^l(k) = \sum_{n=0}^{\infty} \sum_{m=-n}^{n} \tilde{a}_{nm}^*(k) H_{nm}^l(k), \tag{3}$$

where $\tilde{a}_{nm}(k) = (-1)^m a_n^{-m}(k)$ is the representation of $a^*(k,\Omega)$ in the SH domain, and $p^l(k)$ denotes the resulting pressure at the left ear. For the right ear, Eq. (3) is computed with the corresponding right ear HRTF in a similar fashion. In the limiting case, evaluation of the sum in Eq. (3) results in a plane wave representation of the BRIR. In practice, however, the functions $p(k,r,\Omega)$, $H^l(k,\Omega)$ and $H^r(k,\Omega)$ are sampled in space with finite resolution, which implies that the infinite series in (3) must be truncated at some order N to avoid introducing any detrimental effects of spatial aliasing.

3. TIMBRE EQUALIZATION

The practical constraint regarding this series truncation motivates a perceptual comparison of BRIRs generated with different truncation orders. Avni et al. [10] showed that truncating (3) not only restricts the spatial resolution of $p^l(k)$, but also affects its frequency content due to the explicit dependency of b_n on kr and the increased truncation error for $kr > N$ [17], with N being the order limit. This direct impact on the resulting timbre may affect perception and obscure psychoacoustic investigations. To compensate for this effect, Villeval [11] suggested an equalization method, which shall be briefly described in this section.

As a first approximation, the transfer function of the human head can be seen as that of a rigid sphere, which provides an analytic means to quantify the frequency related effects of constructing an incident wave with a finite series of spherical harmonics. In a reverberant setting, the sound field is comprised of a large number of plane waves of random incidence directions. Assuming that a receiver, representing the sound pressure at the ear, is placed at some point Ω_0 on the surface of a rigid sphere, then the average magnitude response over all incident waves is given by [11]

$$\overline{p(kr,\Omega_0)} = \sqrt{\frac{1}{4\pi} \sum_{n=0}^{N} \sum_{m=-n}^{n} |b_n(kr)|^2 \, |Y_n^m(\Omega_0)|^2}, \tag{4}$$

which reduces to

$$\overline{p(kr)} = \frac{1}{4\pi} \sqrt{\sum_{n=0}^{N} |b_n(kr)|^2 \, (2n+1)}, \tag{5}$$

where $b_n(kr)$ is as defined before, and $Y_n^m(\cdot)$ are the spherical harmonics [15]. Accordingly, one can describe the transfer function of a *timbre correction filter*, that equalizes the frequency response of some finite series of order N to that of an order N_h, as follows:

$$H(k)\big|_{N \to N_h} = \frac{\overline{p(kr)}_{N_h}}{\overline{p(kr)}_N}. \tag{6}$$

For example, the magnitude response of two timbre correction filters, equalizing orders $N = 3 \to 19$ and $N = 2 \to 10$, are shown in Figure 1.

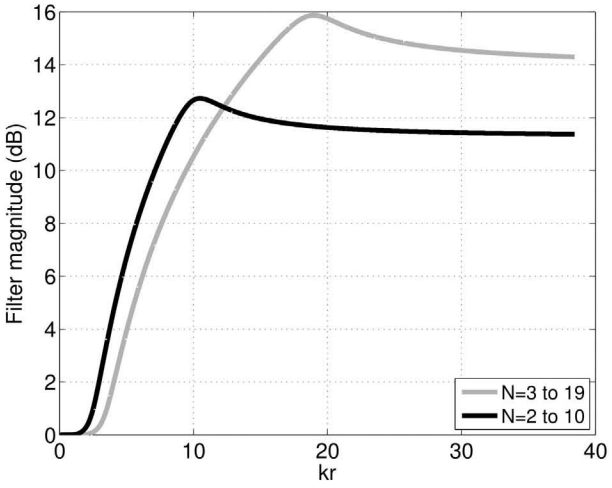


Figure 1: *Magnitude response of two timbre correction filters, equalizing orders* 3 *to* 19 *and* 2 *to* 10.

4. LISTENING TEST

Following the proposed timbre correction filters, the objective of the listening test is to investigate the filters' effects on three perceptual attributes, namely:

1. *Externalization*. When binaurally reproducing a sound field over headphones, the listener may or may not experience a sense of externalization and, accordingly, judges whether the sound is arriving from within the headphones or from a more distant location.
2. *Localization blur / focus*. Whether the sound is externalized or not, it is spatially localized with some error. Thus, a source perceived as having a well defined position in space is here referred to as having a low localization blur or, a high localization focus.

3. *Timbre*. As discussed in Sec. 3, limiting the order of Equation (2) has an effect on the frequency bandwidth. Accordingly, the perceived timbre of the recorded material is modified.

4.1. Methodology

A binaural representation of a sound field was rendered using the method described in Sec. 2, by making use of a pre-measured HRTF set and a SRIR measured using a spherical microphone array. The HRTF set consists of measurements of a Neumann KU-100 manikin, based on a Gaussian sampling scheme with a total resolution of 16020 measurement points distributed around the manikin with no spatial gaps. All technical details regarding the used HRIRs, including the measurement procedure and post-processing can be found in [18]. The chosen SRIR was of the WDR small broadcast studio, having a floor area of 201m^2 and a total volume of 1246m^3 [19], and was sampled using a 1202 points nearly-uniform scheme. Both the HRTFs and the SRIR can be found on-line as part of the the WDR impulse response compilation [20].

To account for head movements (which are required for achieving an effective sense of externalization), BRIR sets were computed for a range of head rotation angles (360° in a 1° resolution), by multiplying $H^l_{nm}(k)$ and $H^r_{nm}(k)$ by respective Wigner-D functions [21]. In the sound reproduction stage, a pair of AKG K702 headphones were fitted with an *Attitude and Heading Reference System* (Razor IMU) which was used to obtain real-time data on the subject's head orientation. All stimuli were processed with a matching headphone compensation filter, were generated pre-test and were played-back using the *SoundScape Renderer* auralization engine [22]. The total latency of the playback system was 5.3ms.

Eleven subjects (all male, ages 24-37) participated in a multiple stimuli, hidden-reference listening test. The labeled reference was based on a BRIR constructed using the method described in Sec. 2, with the SH series truncated at $N = 19$. Similarly, the remaining test samples were based on a BRIR constructed at $N = 3$ (also serving as a low-anchor) and the same BRIR equalized to $N = 19$ using the method described in Sec. 3. To obtain the final test samples, these BRIRs were convolved with anechoic recordings of a classical guitar and of speech. In each screen listeners were asked to rank, on a scale of 1 to 5, how similar each test sample is to the reference in terms of externalization, focus and tone (timbre).

Figure 2 shows the sixth-octave smoothed spectrum of the anechoic speech recording used in the listening experiment. To demonstrate the effects of timbre equalization, the anechoic signal was convolved with three left-ear BRIRs based on $N = 19$ (reference curve), $N = 3$ and $N = 3/E19$ (corresponding to order 3 equalized to order 19). Observe that up to $kr \approx 3$ (equivalent to $f = 1.9\text{kHz}$ for a sphere of $r = 85\text{mm}$), which represents the frequency range of the filter's stop-band, the three curves are nearly identical. Above $kr = 3$, which represents the filter's pass-band, the $N = 3/E19$ curve is amplified compared to the $N = 3$ curve. Because the filter is designed based on the pressure magnitude averaged over all incident directions, then for any room other than a perfectly diffuse field, the accuracy of compensation will always be dependent on the specific directional characteristics of the SRIR.

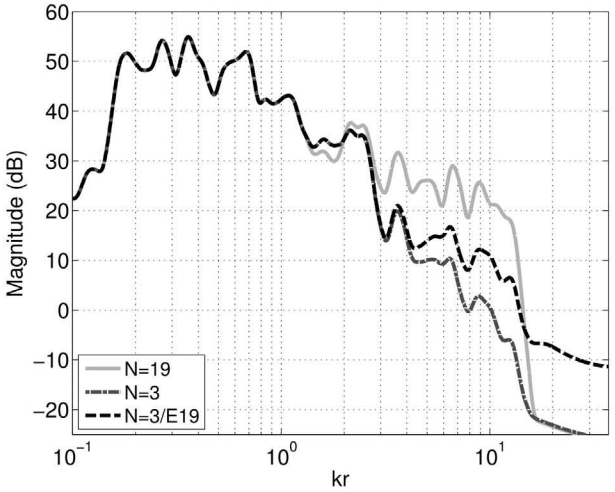


Figure 2: *Sixth-octave smoothed spectra for the anechoic speech recording, after convolving it with left-ear BRIR based on $N = 19$, $N = 3$ and $N = 3/E19$.*

4.2. Results

Figure 3 shows the mean scores ($\bar{x}$) and confidence intervals ($t_{.95,11} = 2.26$) for the different test samples used in the experiment. All eleven subjects gave the hidden reference a score of "5", and as such, this listening condition is excluded from the results.

For the case of timbre, there is a significant difference between equalized and unequalized test samples, for both speech ($\bar{x} = 2.27, \sigma = 0.786$ compared to $\bar{x} = 1.09, \sigma = 0.301$) and classical guitar ($\bar{x} = 2.45, \sigma = 0.82$ compared to $\bar{x} = 1.36, \sigma = 0.67$).

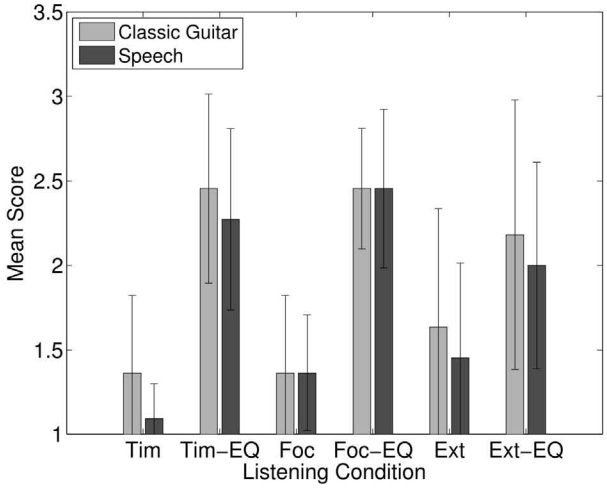


Figure 3: *Mean score and 95% confidence intervals for subjective evaluation of timbre (Tim), focus (Foc) and externalization (Ext), testing for speech (dark bars) and classical guitar (light bars). Equalized BRIRs are marked with "EQ".*

A similar trend is evident in results of the focus test ($\bar{x} = 2.45, \sigma = 0.68$ compared to $\bar{x} = 1.36, \sigma = 0.50$ for speech, and $\bar{x} = 2.45, \sigma = 0.52$ compared to $\bar{x} = 1.36, \sigma = 0.67$ for classical

guitar), suggesting a significant spatial improvement.

In the externalization test, however, the improvement in the mean score is smaller (a difference of only $\Delta\bar{x} = 0.54$ points for both classical guitar and speech) and the confidence intervals overlap. While this does not necessarily indicate a lack of statistical significance (the data is dependent), a more rigorous testing is required to investigate this effect.

5. CONCLUDING REMARKS

A preliminary study on the effects of timbre correction filters on low-resolution BRIR rendering was presented in this paper. A perceptual comparison of equalized vs. unequalized samples revealed a significant improvement in the perceived timbre and a reduction of localization blur. Such localization errors are known to decrease as the order of the microphone array is increased [23]. This immediately suggests that spectral equalization of low order array recordings could be beneficial not only for timbre restoration, but also for improving the perceived spatial resolution, most noticeably in terms of sound localization. One possible reason for this result may be related to the recovery of high-frequency auditory cues, which contribute to sound localization [24] as well as the spatial perception of an enclosed space [25].

Future tests will involve a more systematic comparison of listening conditions, taking into account a wider range of truncation orders, room characteristics, source positions and program-materials.

6. ACKNOWLEDGMENTS

The authors would like to thank Barak Ben-Dayan for his technical assistance in implementing the head-tracking system.

7. REFERENCES

[1] Henrik Møller, "Fundamentals of binaural technology," *Applied acoustics*, vol. 36, no. 3, pp. 171–218, 1992.

[2] Frederic L Wightman and Doris J Kistler, "Headphone simulation of free-field listening. i: Stimulus synthesis," *The Journal of the Acoustical Society of America*, vol. 85, pp. 858, 1989.

[3] Simon Carlile, *Virtual auditory space: Generation and applications*, RG Landes New York, 1996.

[4] Hilmar Lehnert and Jens Blauert, "Principles of binaural room simulation," *Applied Acoustics*, vol. 36, no. 3, pp. 259–291, 1992.

[5] Mendel Kleiner, Bengt-Inge Dalenbäck, and Peter Svensson, "Auralization-an overview," *Journal of the Audio Engineering Society*, vol. 41, no. 11, pp. 861–875, 1993.

[6] Nathaniel I Durlach, A Rigopulos, XD Pang, WS Woods, A Kulkarni, HS Colburn, and EM Wenzel, "On the externalization of auditory images," *Presence: Teleoperators and Virtual Environments*, vol. 1, no. 2, pp. 251–257, 1992.

[7] Durand R Begault, Elizabeth M Wenzel, and Mark R Anderson, "Direct comparison of the impact of head tracking, reverberation, and individualized head-related transfer functions on the spatial perception of a virtual speech source," *Journal of the Audio Engineering Society*, vol. 49, no. 10, pp. 904–916, 2001.

[8] Barbara G Shinn-Cunningham, Norbert Kopco, and Tara J Martin, "Localizing nearby sound sources in a classroom: Binaural room impulse responses," *The Journal of the Acoustical Society of America*, vol. 117, pp. 3100, 2005.

[9] Boaz Rafaely and Amir Avni, "Interaural cross correlation in a sound field represented by spherical harmonics," *The Journal of the Acoustical Society of America*, vol. 127, pp. 823, 2010.

[10] Amir Avni, Jens Ahrens, Matthias Geier, Sascha Spors, Hagen Wierstorf, and Boaz Rafaely, "Spatial perception of sound fields recorded by spherical microphone arrays with varying spatial resolution," *The Journal of the Acoustical Society of America*, vol. 133, pp. 2711, 2013.

[11] Shahar Villeval, "Spatial soundfield processing for headphone reproduction," M.S. thesis, Ben-Gurion University of the Negev, To be published in 2014.

[12] George Brown Arfken, Hans-Jurgen Weber, and Lawrence Ruby, *Mathematical methods for physicists*, vol. 6, Academic press New York, 1985.

[13] James R Driscoll and Dennis M Healy, "Computing fourier transforms and convolutions on the 2-sphere," *Advances in applied mathematics*, vol. 15, no. 2, pp. 202–250, 1994.

[14] Boaz Rafaely, "Plane-wave decomposition of the sound field on a sphere by spherical convolution," *The Journal of the Acoustical Society of America*, vol. 116, pp. 2149, 2004.

[15] Earl G Williams, *Fourier acoustics: sound radiation and nearfield acoustical holography*, Access Online via Elsevier, 1999.

[16] Benjamin Bernschütz, Christoph Pörschmann, Sascha Spors, Stefan Weinzierl, and Begrenzung der Verstärkung, "Soft-limiting der modalen amplitudenverstärkung bei sphärischen mikrofonarrays im plane wave decomposition verfahren," *Proceedings of the 37. Deutsche Jahrestagung für Akustik (DAGA 2011)*, pp. 661–662, 2011.

[17] Munhum Park and Boaz Rafaely, "Sound-field analysis by plane-wave decomposition using spherical microphone array," *The Journal of the Acoustical Society of America*, vol. 118, pp. 3094, 2005.

[18] Benjamin Bernschütz, "A spherical far field hrir/hrtf compilation of the neumann ku 100," *Proceedings of the 40th Italian (AIA) Annual Conference on Acoustics and the 39th German Annual Conference on Acoustics (DAGA) Conference on Acoustics*, 2013.

[19] B. Bernschütz P. Stade and M. Ruhl, "Spatial audio impulse response compilation captured at the wdr broadcast studios," *Proc. TONMEISTERTAGUNG - VDT International Convention (TMT) Cologne, Germany*, 2012.

[20] Benjamin Bernschütz, "Spatial audio impulse response compilation captured at the wdr broadcast studios," `http://www.audiogroup.web.fh-koeln.de/wdr_irc.html`, 2013.

[21] Boaz Rafaely and Maor Kleider, "Spherical microphone array beam steering using wigner-d weighting," *Signal Processing Letters, IEEE*, vol. 15, pp. 417–420, 2008.

[22] Jens Ahrens, Matthias Geier, and Sascha Spors, "The soundscape renderer: A unified spatial audio reproduction framework for arbitrary rendering methods," in *Audio Engineering Society Convention 124*. Audio Engineering Society, 2008.

[23] Stéphanie Bertet, Jérôme Daniel, Etienne Parizet, and Olivier Warusfel, "Investigation on localisation accuracy for first and higher order ambisonics reproduced sound sources," *Acta Acustica united with Acustica*, vol. 99, no. 4, pp. 642–657, 2013.

[24] John C Middlebrooks and David M Green, "Sound localization by human listeners," *Annual review of psychology*, vol. 42, no. 1, pp. 135–159, 1991.

[25] Barbara G Shinn-Cunningham and Suraj Ram, "Identifying where you are in a room: Sensitivity to room acoustics," in *Proc. Int. Conf. Auditory Displays*, 2003, pp. 21–24.

PERCEPTUAL AND ROOM ACOUSTICAL EVALUATION OF A COMPUTATIONAL EFFICIENT BINAURAL ROOM IMPULSE RESPONSE SIMULATION METHOD

Torben Wendt,

Acoustics Group, Medical Physics
and Cluster of Excellence "Hearing4all",
University Oldenburg
Oldenburg, Germany
t.wendt@uni-oldenburg.de

Steven van de Par,

Acoustics Group and Cluster of Excellence "Hearing4all",
University Oldenburg
Oldenburg, Germany
steven.van.de.par@uni-oldenburg.de

Stephan D. Ewert,

Medical Physics and Cluster of Excellence "Hearing4all",
University Oldenburg
Oldenburg, Germany
stephan.ewert@uni-oldenburg.de

ABSTRACT

A fast and perceptively plausible method for synthesizing binaural room impulse responses (BRIR) is presented. The method is principally suited for application in dynamic and interactive evaluation environments (e. g., for hearing aid development), psychophysics with adaptively changing room reverberation, or simulation and computer games. In order to achieve a low computational cost, the proposed method is based on a hybrid approach. Using the image source model (ISM; Allen and Berkley [J. Acoust. Soc. Am. Vol. 66(4), 1979]), early reflections are computed in a geometrically exact way, taking into account source and listener positions as well as wall absorption and room geometry approximated by a "shoebox". The ISM is restricted to a low order and the reverberant tail is generated by a feedback delay network (FDN; Jot and Chaigne [Proc. 90th AES Conv., 1991]), which offers the advantages of a low computational complexity on the one hand and an explicit control of the frequency dependent decay characteristics on the other hand. The FDN approach was extended, taking spatial room properties into account such as room dimensions and different absorption characteristics of the walls. Moreover, the listener orientation and position in the room is considered to achieve a realistic spatial reverberant field.

Technical and subjective evaluations were performed by comparing measured and synthesized BRIRs for various rooms. Mostly, a high accuracy both for some common room acoustical parameters and subjective sound properties was found. In addition, an analysis will be presented of several methods to include room geometry in the FDN.

1. INTRODUCTION

Room acoustical simulations are desirable for many purposes, such as developing or testing signal processing algorithms, or to e. g. test the effect of reverberation on speech intelligibility. Furthermore, they are of interest for audio-visual simulation environments (e. g, for training and rehabilitation) and in entertainment, e. g. in computer games, all requiring a real-time adaptation of the virtual environment, depending on the movement of the listener and/or the sound sources.

One traditional way to emulate the acoustics of a certain room is to measure binaural room impulse responses (BRIRs) and to convolve dry source signals with the BRIRs. However, such measurements are time consuming and their usage is restricted to static scenarios. Furthermore, one is restricted to actually existing rooms. Alternatively, room acoustics can be simulated, enabling different degrees of realism, ranging from simple artificial reverb generation to complex room acoustical simulation (image source model [1], CATT [2], ODEON [3]), even for dynamic scenarios (e. g. [4], [5], [6]).

Depending on the application, physically correct rendering of a soundfield is required or a perceptually convincing auralization, implying plausibility and authenticity, is sufficient. For room simulations used in psychoacoustic research, rehabilitation or in computer games perceptual aspects are most important, implying accordance of room acoustical parameters, e. g. reverberation time, definition, and measures like speech intelligibility. In this case simplifications can be made to reach computational efficiency allowing for real-time rendering of dynamic acoustic scenes, in which the positions of sources and receivers can be changed interactively.

Several approaches exist to synthesize room impulse responses. If the wavelength of a sound is small compared to the characteristic dimensions of reflecting objects, concepts of geometric acoustics (GA), such as the image source model [1] or the ray tracing method [7] can be applied. Both methods have been used and further developed in various room acoustics simulation algorithms, mostly as hybrids together with other algorithms (e. g. [3], [8]). However, these methods still have high computational complexities.

If the exact room geometry is neglected, artificial reverberation can be synthesized very efficiently and with predefined reverberation time. Here, a common approach are feedback delay networks (FDNs), based on Schroeder's pioneering work on parallel delay lines with feedback [9] and further developed (amongst others) by Stautner and Puckette [10] and Jot and Chaigne [11].

One way to achieve real-time performance while maintaining

the advantages of the more "accurate" GA-based BRIR synthesis and reverberation algorithms is their combination in a hybrid approach: The initial part of the impulse responses is computed based on a GA method. The reverberant tail is generated by a more effective reverberation algorithm. Perception motivates such an approach as the early sound reflections create the impression of a certain spatial source width on the one hand and support speech intelligibility on the other hand. The following reverberant tail contains diffuse reflections and its frequency dependent decay characteristics conveys information about the wall absorption and room size.

Here, a hybrid approach was evaluated which combines the image source model (ISM) for a shoebox geometry to simulate early reflections up to a low order, and an FDN for creating a diffuse reverberation tail. The FDN was extended to be directly linked to the room geometry used in the ISM, and to be able to spatially render the reverberation tail in order to generate BRIRs. For low ISM orders, BRIRs can be simulated very efficiently with this approach. In technical and subjective evaluations, the ability of the algorithm to create plausible and authentic simulations was assessed for single and connected (coupled) shoebox rooms. Two different approaches for spatial reverb distribution rendering were compared, taking room dimensions and receiver position into accout.

2. SIMULATION METHOD

A hybrid approach [12] was used to synthesize BRIRs. Early sound reflections are computed by an image source model up to a low order. The late reverberation was generated by a feedback delay network.

The auralization steps are described explicitly for the case of headphone presentation, reflected by the application of head-related impulse responses (HRIRs). The adaptation to arbitrary loudspeaker-based playback systems, such as higher order ambisonics or wave field synthesis, can be easily achieved by replacing them by respective loudspeaker-controlling functions.

2.1. Image source model

The ISM regards a sound reflection as the direct sound of a mirrored version of the original source. This so-called image source differs from the original source by its time delay and its attenuation due to the distance to the receiver, as well as the respective wall reflection coefficient. The sound of an image source is reflected again at other walls, creating higher order image sources. In this way, arbitrarily complex reflection paths can be modeled.

The ISM implementation in the proposed simulation method is restricted to empty shoebox-shaped rooms, where the six wall surfaces are represented each by frequency dependent absorption coefficients. These shoebox-shaped rooms enable a very efficient calculation of image source positions in comparison to arbitrary room geometries [13]. Nevertheless, for a shoebox room the number of image sources up to reflection order N is of order $\mathcal{O}(N^3)$, which considerably affects computational efficiency for higher reflection orders. Another limitation of the ISM is that it inherently assumes only specular instead of diffuse reflections, although they are of importance to describe room acoustics.

In the ISM implementation, the following signal processing steps are performed for each image source: A "$1/$distance" attenuation factor and a time delay due to distance to receiver; an "effective reflection filter", being the (frequency domain) product of all wall reflection coefficients that are involved to "create" the current image source; an HRIR, according to the azimuth and elevation position of the image source relative to the receiver's head orientation. Finally, the binaural signals for all image sources are added up to one two-channel output.

2.2. Extended feedback delay network

The extended FDN used here is based on the general multichannel network as suggested by Jot and Chaigne [11] and consists basically of a set of parallel delay lines whose outputs are fed back via a feedback matrix A.

The number of parallel channels (delay lines) was set to 12, with four channels associated to each (showbox) room dimension (two channels per wall) reflected in several parameter choices. Firstly, the delay units $\tau_j, j \in \mathbb{N}_{\leq 12}$ were directly related to the room dimensions via sound propagation speed (plus a random jitter per channel). Secondly, the absorption filters with transfer functions H_j^{abs} simulate the frequency dependent sound attenuation due to the wall reflections and air absorption. After Jot and Chaigne [11] the frequency dependent reverberation time $T_{60}(f)$ conveyed by the resulting RIR is controlled explicitly by the following frequency responses, if all other processing steps are energy preserving:

$$20 \lg |H_j^{\text{abs}}(f)| = -60\tau_j / T_{60}(f)\,. \tag{1}$$

In the simulation method, the reverberation time is predicted from the wall absorption coefficients via Sabine's formula. Thirdly, the feedback matrix A redistributes the outputs back to the input channels. This process is energy preserving if A is an orthogonal matrix. Here, a randomly created unitary matrix was chosen, providing a high variety of pulse amplitudes.

Two last processing steps per channel, referred to as "binauralization steps", extend the FDN to introduce spatiality distributed and externalized reverberation. (1) Via HRIR filtering the FDN channels are mapped to 12 points (directions) around the head, with two points positioned on each wall. (2) Reflection filters—identical to those applied to the first order image sources in the ISM—simulate a direction dependent sound intensity of reverberation, due to the different acoustical wall properties.

Two possible principles are suggested to map the 12 directions around the receiver's head, which are sketched in Fig. 1 (microphone symbol: recveiver, big "⊗": direct sound source, small "⊗": reverb source). The first one (lhs of Fig. 1) is called "cube" condition. Here, the 12 directions are mapped to points on a cube around the receiver's head. The cube always moves with the receiver (receiver is always in its centre) and is axis aligned with the room.

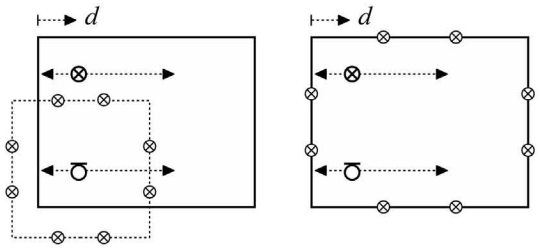


Figure 1: *Illustration of two possible techniques of spatial reverb distribution: "cube" (lhs) and "box" (rhs) method. See text for explanation. (Arrows will be explained in sec. 3.2.*

In this way all 12 incidence directions are more or less equally distributed around the head. In the second "box" condition the 12 directions are mapped to points on the actual six wall surfaces like depicted in the rhs part of Fig. 1. Here, the sound incidences are warped according to the room dimensions and the actual receiver position. Differences between both methods will clearly be audible for rooms with large differences in dimensions. If not specified, in the following the cube condition will be used as standard rendering method.

2.3. Combination of ISM and FDN

For a smooth transition from the early-reflections part (ISM) to the late-reverberation part (FDN), i. e. a straight decay of the BRIR on dB scale, the energy and initial delay of the FDN input signal have to be suitable. For this purpose the FDN input signal consists of the Nth-order ISM pulses before HRIR filtering. In order to avoid comb-filter coloration effects, which occur if a fixed temporal pattern of pulses is fed into the FDN, the ISM output is distributed to the FDN channels. Because the number of image sources of order N does in general not equal the number M of FDN channels, the ith ISM pulse is fed into the FDN channel $[(i-1) \mathrm{mod} M] + 1$.

2.4. Simplified model for coupled rooms

In addition to the single shoebox-shaped room a strongly simplified method to simulate the acoustics of two connected shoebox rooms that are acoustically coupled, e. g. by an open door is suggested.

It is assumed that a source S is located in room 1 and a receiver R in room 2 as depicted in Fig. 2. The sound transmission from room 1 to room 2 is then simulated by a single virtual source S′ located in the door which is exciting room 2 as in the case of the single shoebox simulation described above. The virtual source radiates the monaural impulse response of room 1 for a source position specified by the coupled-room arrangement and a "monaural" receiver R′ inside the open door. Thus, the effective BRIR is obtained as the convolution product of the monaural RIR of room 1 with the BRIR of room 2.

Depending on the source position in room 1, it is either visible or invisible for a receiver in room 2. If it is not visible, no direct sound will arrive at the receiver but only reflections and diffractions. In this case, the direct sound pulse of the RIR of room 1 is discarded in the current approach.

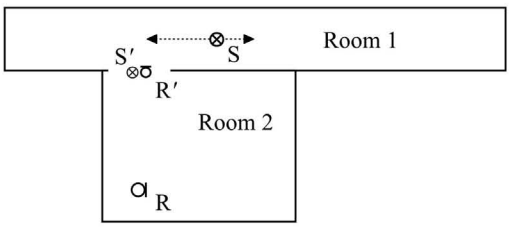


Figure 2: *Sketch of two adjacent rooms, that are acousticly coupled by an open door. See text for explanation. (The arrow will be explained in sec. 3.2.)*

3. EVALUATIONS

Three main aspects of the proposed simulation method were evaluated. Firstly, for a set of real-existing rooms, subjectively rated sound properties of measured and respectively synthesized BRIRs were compared. Secondly, the two approaches to realize the binauralization steps of the extended FDN (see sec. 2.2) were evaluated with respect to binaural parameters. Thirdly, the approach to simulate the acoustics of two coupled rooms was evaluated.

To perform these evaluations, a test-database containing measured and synthesized BRIRs was created. BRIRs were measured for various rooms of different size and reverberation time, as well as a for a few source-receiver configurations in two connected rooms. Additionally, some measured BRIRs were taken from the AIR database [14].

The BRIR measurements were performed using an omnidirectional loudspeaker based on a ring-radiator and an artificial head MK2 by Cortex. Rooms were excited with a logarithmic sweep [15] (50 Hz to 18 kHz) offering removal of nonlinear harmonic distortions from the recorded and inverse filtered signal. BRIRs were each calculated as the mean of BRIRs from 10 single recordings and equalized by the inverse loudspeaker transfer function.

For the BRIR synthesis a single mean wall absorption coefficient was used. It was determined for each room from its reverberation time via the inverse form of Sabine's formula, which ensures that reverberation times of measured and synthesized BRIRs are in good accordance.

The HRIRs used in the simulation were from the same artificial head as used for the BRIR recordings. The database [16] offers HRIRs with azimuth angles in 2° steps at elevations near the equatorial level. Towards the poles, the azimuth angle sampling decreases. Elevation angles are sampled in 2° steps.

A varying synthesis parameter was the maximum image source order. The goal was to find a trade-off between accuracy and computational efficiency.

In the following, for all different rooms and source-receiver configurations, the term "room condition" will be used. In contrast, different types of BRIR synthesis, differing in the choice of simulation parameters, will be referred to as "synthesis conditions". All room- and synthesis conditions will be introduced in the following.

3.1. Subjective sound properties

3.1.1. Room- and synthesis conditions

BRIRs were chosen from four rooms of different size and reverberation time, specified in Tab. 1. For the synthesized BRIRs, the maximum image source order N was varied in $\{1, 3\}$. For one room the BRIR was synthesized only by the ISM with $N = 20$. In the extended FDN, the cube condition was chosen (see 2.2). Two dry source signals, female spoken speech and a guitar play (steel strings) were convolved with the recorded and simulated BRIRs. Presentation sound pressure levels ranged from 60 to 65 dB SPL, depending on the source-receiver distance and the room reverberation.

3.1.2. Subjects and procedure

15 normal-hearing subjects (7 female, 8 male) aged 24 to 32 years participated in the experiment. Sounds were presented via headphones in a sound attenuating booth. Since the synthesis method was implemented as an offline simulation, no head tracking and adaptively changing soundfield was employed.

The sound properties which were to be rated on a seven-point scale were "naturalness" and "room size". A test and a retest were

Table 1: *Rooms, whose BRIRs were used in the subjective evaluation. Reverberation times T_{60} were obtained from measured BRIR (broadband).*

Room	Dimensions (m)	T_{60} (s)
Aula	(12.0, 30.0, 10.0)	4.8
Empty chamber	(1.88, 2.74, 2.82)	2.5
Lecture room	(10.90, 10.80, 3.15)	0.8
Laboratory	(4.97, 4.12, 3.00)	0.3

performed in two sessions, each with a randomized order of presented sounds. Before the actual experiment was performed, sound examples illustrating extremal distinctions of the sound properties had been presented.

3.1.3. Results

Fig. 3 shows the results from the subjective sound property ratings as mean values over all subjects and source signals. Each panel shows results for one sound property. The avarage ratings are plotted for all synthesis conditions against rooms. Error bars indicate inter-subject standard errors.

For naturalness (left panel) ratings differ strongly between rooms. Whereas the BRIRs of the laboratory were rated to sound most natural, lowest naturalness was perceived for the empty chamber. BRIRs of the aula and lecture room were rated to have a medium to high naturalness. Between synthesis conditions, almost no differences are visible, and for most rooms, differences between synthesized and measured BRIRs are very low. Moreover, for some conditions even the synthesized BRIRs were rated to sound slightly more natural than the measured one. This shows, that the proposed simulation method is able to synthesize BRIRs that sound as natural as measured ones. Remaining differences in perceived naturalness between rooms might be due to familiarities of subjects with these acoustic environments in daily life, since the Laboratory sounds as dry as an ordinary living room, whereas the empty chamber sounds rather unusual, even by the measured BRIR. This might also be due to the unusual relation of its very small room size and its high reverberation time (see Tab. 1).

For room size (right panel), again clear differences between rooms are perceived. The order is well in accordance with reverberation times and, except for the empty chamber, with the actual room sizes (see Tab. 1). Differences within synthesis conditions and between synthesises and measurements are practically not existent.

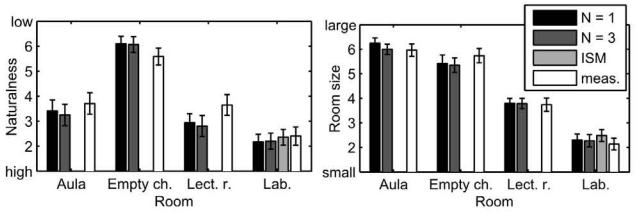


Figure 3: *Subjective sound property ratings of measured and synthesized BRIRs for four rooms, averaged over all subjects and source signals (speech and music). Error bars indicate inter-subject standard errors.)*

This shows that firstly the simulation method is able to represent different room sizes and secondly to achieve this independently from maximum image source order as far as tested.

As a consequence of the experimental design, no direct mapping of synthesized BRIRs to actual rooms was performed. Given that no head tracking was employed, a potential effect of head rotations on the subjective ratings could so far not be assesed. Future research will apply the system in a real-time environment and will address this issue.

3.2. Evaluation of spatial properties of the extended FDN

3.2.1. Room- and synthesis conditions

Two rooms each with different configurations of source- and receiver positions as well as wall properties, were used to evaluate the spatial reverb rendering. Fig. 1 depicts schematically the conditions for one room. Tab. 2 specifies the room dimensions and the absorption coefficients for (250, 500, 1k, 2k, 4k) Hz. While room 1 has an almost square base area, room 2 represents a long corridor. Side wall absorption coefficients were specified in two different ways: In the "closed" condition, all side wall absorption coefficients were equal as given in line 3 in Tab. 2. By this way, spatial sound properties in azimuth can be investigated in dependence of room geometry in connection to the positions of the receiver and all virtual "reverb" sources. In the "open" condition, the left side was completely open, meaning that no wall was existent. This was technically represented by a broadband absorption coefficient of 0.99, whereas the absorption of all other side walls did not differ from those of the closed condition. By this, the spatial rendering was to be evaluated in a challenging condition for the model.

Also in both rooms and for the closed- and open condition, the distance d of source and receiver to the left wall was chosen to be 0.3 m or 5 m (see Fig. 1). The source was always in the front direction of the receiver, yielding a direct sound with no interaural differences. All differences are thus due to reflections and reverberation.

Since no suitable real rooms were found for BRIR measurements, the ISM with reflection order 20 was used as reference. For all room conditions, the cube- and box condition were compared against each other and the reference. ILDs and IACCs were determined from the BRIR up to the time $\min\{T_{60}(f)\}$, where $T_{60}(f)$ is the frequency dependent reverberation time. Positive ILDs indicate higher signal energy on the right.

Table 2: *Specification of two virtual rooms used for evaluation of the binauralization steps in the extended FDN. See text for further explanation.*

Dimensions room 1	(10.9, 10.8, 3.15) m
Dimensions room 2	(3.9, 30.0, 3.15) m
Absorption side walls	(0.05, 0.10, 0.13, 0.16, 0,22)
Absorption open side	0.99
Absorption floor	(0.03, 0.03, 0.03, 0.03, 0.02)
Absorption ceiling	(0.70, 0.60, 0.70, 0.70, 0.50)

Table 3: *Results of the evaluation of the FDN binauralization steps: Comparison of cube and box method with purely ISM-created BRIRs in terms of ILDs and IACCs. See text for explanation of conditions.*

	ISM	cube	box	ISM	cube	box
ILD (dB)				IACC		
R 1, closed						
$d = 0.3$ m	-0.3	$\mathbf{-0.7}$	1.3	0.5	0.5	0.5
$d = 5.0$ m	0.3	**0.7**	0.8	0.7	0.7	0.7
R 1, open						
$d = 0.3$ m	0.9	**0.9**	1.8	0.9	0.8	0.8
$d = 5.0$ m	1.2	1.1	**1.2**	0.8	0.8	0.8
R 2, closed						
$d = 0.3$ m	-0.6	$\mathbf{-0.2}$	3.1	0.6	**0.5**	0.4
$d = 5.0$ m	-0.6	**0.7**	3.7	0.7	**0.6**	0.5
R 2, open						
$d = 0.3$ m	0.5	**1.0**	3.2	0.9	**0.8**	0.6
$d = 5.0$ m	0.5	**1.0**	3.0	0.9	**0.8**	0.6

3.2.2. Results

The results for all room- and synthesis conditions are shown in Tab. 3. Comparing results for the cube and box condition, values that are closer to those of the ISM reference, with a difference of at least 0.1 dB (ILD) or 0.1 (IACC) to the less matching condition, are printed in bold face for clarity. To interpret the results just-noticeable differences (JNDs) of ILD have to be considered as, e. g. determined in [17] for musical instruments in several reverberant conditions: 1.0–1.4 dB for $T_{60} = 1.3$ s; 0.8–1.2 dB for $T_{60} = 0.8$ s; 0.4–0.8 dB for the anechoic condition.

For room 1 ("R 1") (closed) overall small absolute ILD values are observed in the range of the JNDs. It has to be kept in mind, that these ILDs originate from reflections only, given that the direct sound was always located in the front direction. For the box condition ILDs are larger and differ considerably more from the reference. Clear mismatches to the reference ILD are obtained in the close-to-wall position ($d = 0.3$ m). Overall small or vanishing ILDs are plausible for the closed conditions since all side walls are equal in absorption coefficient. For the closed room 2 ("R 2"), both ISM and cube condition show again small absolute ILD values. However, ILDs for the box condition differ clearly from those of the cube and reference conditions. This is not surprising because the majority of virtual reverb sources lie clearly to the right hand side of the receiver (see also scheme in Fig. 1). For the open versions of the rooms, ILDs obtained from ISM-created BRIRs and the cube condition have very similar values. Largest differences are again obtained for the box condition in room 2.

The IACC results reveal overall no distinct differences between the synthesis conditions for room 1. For room 2, where maximum differences are 0.3, the cube condition yields IACCs that are closer to those created by the ISM.

In conclusion, it can be said that the cube condition mostly creates spatially more realistic BRIRs in terms of ILD and IACC than the box condition, especially when the room geometry and receiver position are challenging. In addition, also an informal subjective listening test yielded highest perceptive similarity between the cube and the reference condition.

3.3. Evaluation of simulation of coupled rooms

3.3.1. Room- and synthesis conditions

The two adjacent rooms, an office and a corridor, acoustically coupled by an open door, are specified in Tab. 4 in terms of dimensions and absorption coefficients for (250, 500, 1k, 2k, 4k) Hz. The arrangement of both rooms and positions of source S and receiver R are depicted in Fig. 2. Two source positions were investigated. In the "visible" condition the source is placed at the left end of the double arrow, and in the "invisible" condition it is placed at the right end. Measured BRIRs for two real rooms from which data in Tab. 4 were obtained served as reference. The ISM condition and the proposed hybrid method with $N = 3$ and $N = 1$ were evaluated. In both hybrid conditions, the "cube" synthesis was used.

Besides a comparison of the BRIRs in the time domain, ILDs and IACCs were determined as described in sec. 3.2.1 and compared with the reference.

Table 4: *Specification of rooms used in the evaluation of the simulation of coupled rooms.*

Room 1 (corridor)	
Dimensions	(30.0, 1.94, 2.50) m
Absorption coeff.	(0.16, 0.16, 0.13, 0.15, 0.17)
Room 2 (office)	
Dimensions	(4.43, 4.50, 3.00) m
Absorption coeff.	(0.25, 0.30, 0.35, 0.32, 0.28)

3.3.2. Results

Fig. 4 shows normalized BRIR time signals for the measured (upper panels) and synthesized ($N = 3$, lower panels) case on an arbitrarily scaled ordinate. As expected for coupled rooms, the measured BRIR in the invisible condition (rhs) shows a rising amplitude in the beginning. This effect can hardly be observed in the simulated BRIR. In this simple approach here, only one convolution of two single RIRs was used which cannot mimick real coupling of the rooms.

Tab. 5 shows ILDs and IACCs obtained from BRIRs of all conditions. For all of them a clear dominance of sound energy on the left is obtained (negative ILDs), which is primarily due to the direction of the (virtual) direct sound (source S$'$ in Fig. 2). The ISM-created BRIRs, which can be assumed to simulate the real rooms best, have indeed ILDs that are closest to those of the measured BRIRs. The ILDs of the hybrid method BRIRs differ maximally 3.4 dB from measurement condition, which is clearly above the JND in reverberant conditions, at least for frontal source positions [17].

For the IACC, all room- and synthesis conditions yield very small values. A slightly higher accordance with the measurement is obtained for the ISM synthesis, but it is questionable, whether these differences were audible.

Concluding, the evaluation showed that this simple approach has limitations if the acoustics of coupled rooms should be simulated in a convincing way. Improvements should consider removal of the direct path between the virtual source and the receiver in the invisible condition. In a second step diffraction of the direct sound

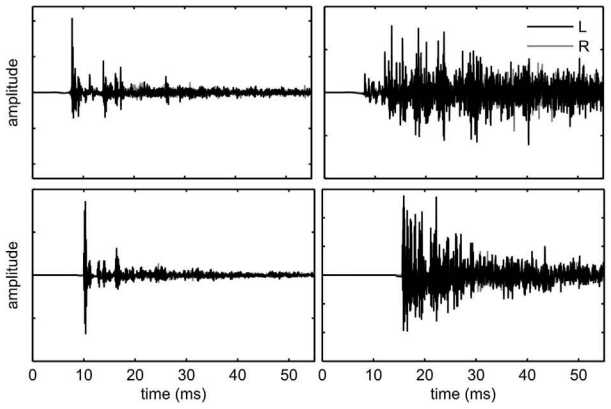


Figure 4: *BRIR time signals on arbitrary amplitude scale. Lhs: visible condition, rhs: invisible condition. Upper panels: measured BRIRs, lower panels: synthesis with* $N = 3$.

Table 5: *Results of the evaluation of the coupled-rooms simulation. Comparison of ILDs and IACCs. See text for explanation of conditions.*

		meas.	ISM	$N = 3$	$N = 1$
ILD (dB)	visible	-6.0	-6.6	-9.4	-8.3
	invisible	-4.5	-6.3	-7.8	-7.8
IACC	visible	0.1	0.1	0.2	0.2
	invisible	0.1	0.1	0.2	0.2

can be taken into account by inclusion of lowpass-filtered versions of the direct path, depending on the geometric relation of the door opening and the source and receiver positions.

4. SUMMARY AND CONCLUSIONS

A hybrid approach for synthesizing binaural room impulse responses of shoebox-shaped rooms was presented. It computes geometrically exact early reflections using the image source model up to a low reflection order, and approximates the reverberant tail by a high efficient feedback delay network. The FDN was extended to enable a spatial reverb rendering, taking into account room geometry as well as wall absorption and source- and receiver positions.

The proposed simulation method was evaluated with respect to different properties using subjective and technical measures. In a subjective evaluation subjects rated the naturalness and room size of measured and respectively synthesized BRIRs. The ratings show that the simulation method is able to represent perceived naturalness and room size very well and independently from maximum image source order, whereas differences in these properties between rooms are clearly conveyed.

For the extended FDN, two spatial reverb rendering techniques (sec. 2.2) were compared in a technical evaluation assessing interaural level differences and interaural cross correlation coefficients. It was shown that synthesized spatial reverberation has better accordance with purely ISM-created reference BRIRs if the reverberation emitting virtual sound sources are equally distributed around the listener's head. In comparison, positioning these sources on the actual room wall surfaces yielded worse results (sec. 3.2.2).

A first, simple approach to simulate the acoustics of two adjacent coupled rooms was evaluated by comparing time signal representations, ILDs and IACCs for measured and synthesized BRIRs. While the results for this approach were not fully convincing future improvement with refined approximations can be expected.

In conclusion, the evaluation showed that the suggested computationally efficient approach for synthesizing binaural room impulse responses is suited for applications where perceptual plausibility and authenticity is acceptable.

5. ACKNOWLEDGEMENTS

This work was supported by the DFG FOR 1732 and the Cluster of Excellence EXC 1077/1 "Hearing4all".

6. REFERENCES

[1] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *J. Acoust. Soc. Am.*, vol. 66, no. 4, pp. 943–950, 1979.

[2] B.-I. Dalenbäck, "Engineering principles and techniques in room acoustics prediction," in *BNAM, Bergen, Norway, May 2010*, 2010.

[3] G. M. Naylor and J. H. Rindel, "Predicting Room Acoustical Behaviour with the ODEON Computer Model," in *124th ASA meeting New Orleans, November 1992*, 1992.

[4] B.-I. Dalenbäck and M. Strömberg, "Real Time Walkthrough Auralization – The First Year," Tech. Rep., CATT (Dalenbäck), Valeo Graphics (Strömberg), 2010.

[5] D. Schröder, F. Wefers, S. Pelzer, D. S. Rausch, M. Vorländer, and T. Kuhlen, "Virtual Reality System at RWTH Aachen University," in *Proceedings ICA 2010, 20th International Congress on Acoustics: 23–27 August 2010, Sydney, New South Wales, Australia*, 2010.

[6] A. Silzle, P. Novo, and H. Strauss, "IKA-SIM: A system to generate auditory virtual environments," in *Audio Engineering Society Convention 116*, 2004.

[7] A. Krokstad, S Strøm, and S. Sørsdal, "Calculating the acoustical room impulse response by the use of a ray tracing technique," *J. Sound Vib.*, vol. 8, no. 1, pp. 118–125, 1968.

[8] Steven M. Schimmel, Martin F. Müller, and Norbert Dillier, "A fast and accurate »shoebox« room acoustics simulator," Tech. Rep., 2009.

[9] M. R. Schroeder, "Natural Sounding Artificial Reverberation," *Journal of the Audio Engineering Society*, vol. 10, no. 3, pp. 219–223, 1962.

[10] J. Stautner and M. Puckette, "Designing Multi-Channel Reverberators," *Computer Music Journal*, vol. 6, no. 1, pp. 52–65, 1982.

[11] J.-M. Jot and A. Chaigne, "Digital delay networks for designing artificial reverberators," in *90th AES Convention*, 1991.

[12] T. Wendt, S. van de Par, and S. D. Ewert, "A computational efficient and perceptually plausible algorithm for binaural room impulse response simulation," *subm. to Journal of the Audio Engineering Society*.

[13] J. Borish, "Extension of the image model to arbitrary polyhedra," *J. Acoust. Soc. Am.*, vol. 75, no. 6, pp. 1827–1836, 1984.

[14] M. Jeub, M. Schäfer, and P. Vary, "A binaural room impulse response database for the evaluation of dereverberation algorithms," Tech. Rep., 2009.

[15] A. Farina, "Simultaneous measurement of impulse response and distortion with a swept-sine technique," in *Audio Engineering Society Convention 108*, 2 2000.

[16] G. Geißler and S. van de Par, "Messung von HRTF am Kunstkopf MK 2 von Cortex," AG Akustik, Carl-von-Ossietzky-Universität Oldenburg, 2012.

[17] S. Klockgether and S. van de Par, "Just Noticable Differences of Spatial Perception in Directly Manipulated Binaural Room Impulse Responses," in *AIA/DAGA 2013, Merano, Italy*, 2013.

EXPLORATION OF A BIOLOGICALLY INSPIRED MODEL FOR SOUND SOURCE LOCALIZATION IN 3D SPACE

Symeon Mattes, *

ISVR Acoustics Group
University of Southampton,
Southampton, UK
symeon.mattes@soton.ac.uk

Philip Arthur Nelson

ISVR Acoustics Group
University of Southampton,
Southampton, UK
p.a.nelson@soton.ac.uk

Filippo Maria Fazi

ISVR Acoustics Group
University of Southampton,
Southampton, UK
ff1@isvr.soton.ac.uk

Michael Capp

Research Department
Meridian Audio Ltd,
Cambridgeshire, UK
Michael.Capp@meridian.co.uk

ABSTRACT

Sound localization in 3D space relies on a variety of auditory cues resulting from the encoding provided by the lower and higher regions of the auditory path. During the last 50 years different theories and models have been developed to describe psychoacoustic phenomena in sound localization inspired by the processing that is undertaken in the human auditory system. In this paper, a biologically inspired model of human sound localization is described and the encoding of the known auditory cues by the model is explored. In particular, the model takes as an input binaural and monaural stationary signals that carry information about the Interaural Time Difference (ITD), the Interaural Level Difference (ILD) and the spectral variation of the Head Related Transfer Function (HRTF). The model processes these cues through a series of linear and nonlinear units, that simulate the peripheral and the pre-processing stages of the auditory system. The encoded cues, which in the model are represented by excitation-inhibition (EI) and the time average (TA) activity patterns, are then decoded by a central processing unit to estimate the final location of the sound source.

1. INTRODUCTION

Sound localization is a perceptual process that in contrast to other sensory systems, like vision and taste, there is no point-to-point correspondence between a sound event and the perceived locus of an acoustic image at the lower peripheral stages of the human hearing system [1]. Instead, it is believed that the localization of sound events occur entirely as a consequence of neural processing of monaural and binaural signals. The ITDs (interaural time differences), the ILDs (interaural level differences), and the monaural spectral cues, that occur due to the spectral changes of the pinna, are three of the most salient auditory cues that are used by a human listener in order to characterize the locus of a sound event.

During the last 50 years different techniques have been developed to predict the statistical properties of human sound localization in the horizontal plane. Some of these theories rely only on stimulus statistics, while others are based on neuroscientific findings. The last one has led to the development of so called biologically inspired models and to three of the most established and well-known theories, i.e. the Jeffress's coincidence detector [2], that is based on coincidence counter hypothesis, Durlach's EC (equalization-cancellation) theory [3], that was developed to interpret phenomena in the detection of binaural sounds masked by a masking noise, and the count-comparison principle introduced by von Békésy (1930) [4] that resembles the neural activity of the higher regions of the auditory path.

At the same time only recently, a variety of different models have been developed for the prediction of human sound localization in sagittal planes [5, 6]. These models are based mainly on the neural integration hypothesis, which states that for moderate intensities the localization system requires an input of at least 80 ms broadband sound to give a stable estimation of the sound-source elevation [7, 8].

Having such models, i.e. a model that is able to predict successfully under certain conditions, human modes of listening, can be beneficial not only for the better understanding of the underlying mechanisms of human reactions but also for their application in audio quality assessment, robotics and cochlear implants, avoiding costly and time-consuming experiments.

The current paper aims to combine two well established models for the prediction of human localization in horizontal and sagittal planes in order to predict human localization in 3D space. The paper is divided into five main sections. In the first section a general introduction to sound localization and to perceptual models is given and in the second section, a biologically inspired model is described for the prediction of human sound localization for stationary signals in 3D space (excluding distance). In the third section, different parameters of the model are explored, and in the third section, simulation results are compared with previous listening tests. In the last part the conclusions and future work are given.

2. DESCRIPTION OF THE MODEL

The model that has been used in this paper is based on EC theory for the production of the excitation-inhibition (EI) pattern in binaural processing [9], which is mainly responsible for the encoding of the ITD and ILD cues, and a time average (TA) representation

* This work was supported by Meridian Audio Ltd.

of a narrow band filtered signal for the production of the monaural processing [6], which is responsible for the encoding of the spectral variations of the HRTFs.

In particular, the model consists of three main stages, each of which corresponds to different (and more or less known) operations of the human auditory system in spatial hearing. The model starts with the peripheral processor, which takes binaural signals as an input. This stage consists of a unit which corresponds to a time-invariant band pass filter from 1 kHz - 4 kHz with a roll-off of 6 dB/octave below 1 kHz and -6 dB/octave above 4 kHz, which represents the response of the human middle ear. This is followed by a fourth-order gammatone filterbank with 100 channels between 100 Hz and 20 kHz [10], which represent the frequency selectivity of the basilar membrane. Each gammatone filter output is processed by a half-wave rectifier, a fifth-order low pass filter with a cut-off frequency at 770 Hz, and a square root compressor, which respectively represents the organ of Corti [11], the gradual loss of the phase-locking in neural firing [12], and the nonlinearities of the basilar membrane in steady state conditions[13].

The model continues with the pre-processor, which consists of one binaural and two monaural units. Each of these units creates three types of patterns ($EI_{k,\tau,\alpha}$, TA_{L_k} and TA_{R_k}) correspondingly, that are compared in the central-processor with a database of patterns by applying a comparison metric which consists of frequency independent functions (m_{bin}, m_L and m_R), called similarity measure (SM) functions [14]. A mapping function is applied to transform m_{bin}, m_L and m_R into the transformed similarity measure function s_{bin}, s_L and s_R. All these functions are then combined to give a single function that represents the likelihood of subject localization of the virtual source.

More specifically, in the pre-processor, the binaural unit, as described by Park et al. [9], is based on the EC theory for the extraction of the excitation-inhibition (EI) cell activity patterns (EI-patterns) and is responsible for the characterization of the position of a lateralized sound source. Given that $L_k(t)$ and $R_k(t)$ are the input signals from the left and the right peripheral processor from the k-th channel of the gammatone filterbank, then each EI unit is characterized by the equation

$$EI_{k,\tau,\alpha}(t) = \left(10^{\frac{\alpha}{40}} L_k(t+\frac{\tau}{2}) - 10^{-\frac{\alpha}{40}} R_k(t-\frac{\tau}{2})\right)^2 \quad (1)$$

where τ is the characteristic ITD in seconds and α the characteristic ILD in dB that occur due to the comparison of the signals of the left and the right ear. At 44.1 kHz sampling frequency the dynamic range is $\pm 700\mu$sec for the characteristic ITD and ± 10 dB for the characteristic ILD, with a resolution of 45 μsec and 1 dB respectively.

Thereafter, the EI-cell activity is normalised by the energy of the input signals associated with a specific snapshot in time, so as to remove any dependency of the amplitude of the input signal. In this case the binaural unit is described by the equation

$$EI''_{k,\tau,\alpha} = \frac{EI'_{k,\tau,\alpha}}{\sqrt{2 e_L e_R}} \quad (2)$$

where e_L and e_R are the energy of the left and the right input signals correspondingly and $EI'_{k,\tau,\alpha}$ is an integrated weighted snapshot over the time t, defined as

$$EI'_{k,\tau,\alpha}(t) = \int EI_{k,\tau,\alpha}(t+t')w(t')\mathrm{d}t' \quad (3)$$

and $w(t)$ is a double-sided exponential window that takes into account the finite binaural temporal nature of the EI-cell activity [9].

The two monaural units are based on the hypothesis that a time average (TA) representation of the narrow band filtered signal that arrives from the peripheral processing unit can be used for the representation of the spectral variations that are necessary for the characterization of an elevated sound source. In this case each unit is characterized by the equation

$$y_k(t) = \frac{1}{T}\int_0^T x_k(t)\mathrm{d}t \quad (4)$$

where $x_k(t)$ is the output of each of the k gammatone filters for the left ($L_k(t)$) and the right ear ($R_k(t)$) integrated over a snapshot of the signal of duration T, which for the current paper the whole duration of the signal has been taken, and $y_k(t)$ is the corresponding monaural pattern for the left (TA_{L_k}) and the right (TA_{R_k}) ear.

The model ends with the central processing unit which is a decision making device that uses a simple pattern matching process in order to characterize the location of the sound source in 3D space. More specifically, the EI-patterns and the TA-patterns that have been produced by a sound source from an unknown location are compared with a bank of EI- and TA-pattern templates in order to produce a SM that quantifies the degree to which the patterns produced by a given source matches the stored patterns.

Given the stationarity and the uniqueness of the sound source, a pattern-matching procedure has been applied for measuring the similarity of the EI-patterns at each channel k of the gammatone filterbank and is defined as

$$\rho_{bin_k}(\phi,\theta) = \frac{\langle EI''_{k,\tau,\alpha},\, EI''_{k,\tau,\alpha}(\phi,\theta)\rangle}{\|EI''_{k,\tau,\alpha}\|\|EI''_{k,\tau,\alpha}(\phi,\theta)\|} \quad (5)$$

where ϕ and θ are the azimuth and elevation angle of the sound source in the interaural-polar coordinate system (fig. 1), $EI''_{k,\tau,\alpha}$ is the EI-patterns of eq. 2 of the target source for a specific azimuth ($\hat{\phi}$) and elevation angle ($\hat{\theta}$), $EI''_{k,\tau,\alpha}(\phi,\theta)$ is the template of the EI-patterns of eq. 2 for all possible azimuth (ϕ) and elevation (θ) positions at the same snapshot, $\langle\cdot\rangle$ is the inner product and $\|\cdot\|$ is the L^2 norm of the EI'' over τ and α.

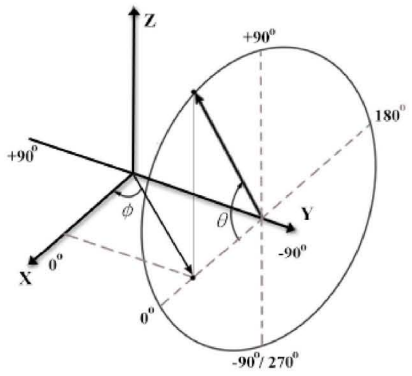


Figure 1: The interaural-polar coordinate system is a head-related spherical coordinate system whereby different azimuth angles ϕ define a cone of confusion. Its range for the azimuth angle is $\phi \in [-\frac{\pi}{2}, \frac{\pi}{2}]$ and for the elevation angle $\theta \in [-\pi, \pi)$ or $\theta \in [-\frac{\pi}{2}, \frac{3\pi}{2})$ [15, 16].

The frequency dependent SM is then weighted in order to give the total SM for the binaural cues, defined as

$$m_{bin}(\phi,\theta) = \sum_k \rho_{bin_k}(\phi,\theta) q_k \tag{6}$$

where q_k is a weighting coefficient that depends on the frequency of the gammatone filter and which varies smoothly with frequency but which reflects the dominance of the binaural cues around 600 Hz [17].

Finally, a mapping function is applied which gives the transformed SM for the binaural cues, defined as

$$s_{bin}(\phi,\theta) = m_{bin}(\phi,\theta)^{\gamma_{bin}} \tag{7}$$

where γ_{bin} modifies the transformed SM, and as demonstrated in sections 3.3 and 4, this will allow comparison with experimental data.

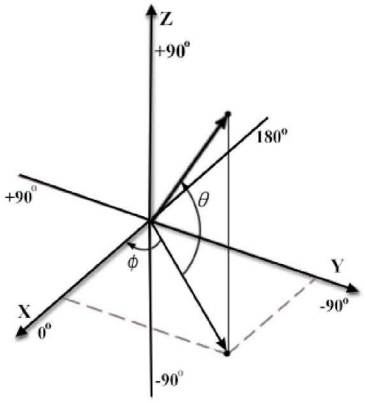


Figure 2: The vertical-polar coordinate system is a head-related coordinate system which is a sub-category of the spherical coordinate system. Its range for the azimuth angle is $\phi \in [-\pi, \pi)$ and for the elevation angle $\theta \in [-\frac{\pi}{2}, \frac{\pi}{2})$ [15].

The SM that has been used for the monaural cues is that suggested by Baumgartner et al. [6] and is the standard deviation of the interspectral differences, defined for the left monaural processor as

$$m_L(\phi,\theta) = \sqrt{\frac{1}{N}\sum_k \left(d_{L_k}(\phi,\theta) - \overline{d_{L_k}(\phi,\theta)}\right)^2} \tag{8}$$

where N is the number of the gammatone filters that has been used in the peripheral processing units, $d_{L_k}(\phi,\theta) = TA_{L_k} - TA_{L_k}(\phi,\theta)$ is the interspectral difference between the TA patterns (TA_{L_k}) of the target source (eq. 4) for a specific azimuth ($\hat{\phi}$) and elevation angle ($\hat{\theta}$) and the template of the TA-patterns ($TA_{L_k}(\phi,\theta)$) of eq. 4 for all available positions in the interaural coordinate system, and $\overline{d_{L_k}(\phi,\theta)}$ is the average value. Similar to eq. 8 , $m_R(\phi,\theta)$ gives the SM for the right monaural preprocessing unit.

Furthermore the SM of the monaural cues are combined through a weighted function as described by

$$s_{mon}(\phi,\theta) = b(\phi)s_L(\phi,\theta) + b(-\phi)s_R(\phi,\theta) \tag{9}$$

where $b(\phi)$ is a weighting function that is based on the assumption that the contralateral ear contributes less to the perception of sound localization than the ipsilateral ear [18], and

$$s_{L/R}(\phi,\theta) = \frac{1}{\sigma_{mon}\sqrt{2\pi}} \mathrm{e}^{-\frac{m_{L/R}(\phi,\theta)}{2\sigma^2_{mon}}} \tag{10}$$

is the mapping function, where $m_{L/R}(\phi,\theta)$ is the SM of the monaural cues for the left ($m_L(\phi,\theta)$) and the right ear ($m_R(\phi,\theta)$), and σ_{mon} again, as shown in sections 3.3 and 4, modifies the mapping function in a way that will allow comparison of the likelihood of localisation with experimental results.

By analogy with the laws of probability we multiply the two transformed SM ($s_{mon}(\phi,\theta)$ and $s_{bin}(\phi,\theta)$), as described by

$$s(\phi,\theta) = s_{bin}(\phi,\theta)s_{mon}(\phi,\theta) \tag{11}$$

to obtain a representation of the likelihood of the subject's localization of the virtual source.

3. EXPLORING THE LOCALISATION CUES

Two of the main characteristics of the model described in sec. 2 are the TA and EI patterns that are constructed through a process that attempts to emulate the human auditory path. These patterns contain information of the static cues associated with the ITD, the ILD and the spectral variations induced by the two pinnae and as a consequence information on the location of a given sound source. The aim of the following sections is to analyze some of the features of the TA and the EI patterns by using the HRTFs of a KEMAR with a small pinna from the CIPIC database (subject 165) [16].

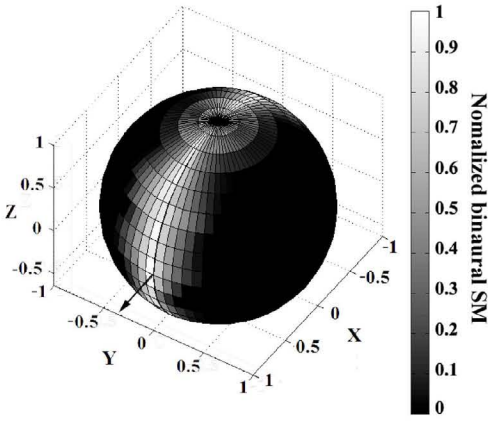


Figure 3: The results of the comparison of the EI patterns at $f = 100$ Hz for a sound source at $\hat{\phi} = \hat{\theta} = 0°$ by using the vertical-polar coordinate system. The colour bar indicates the value of eq. 5 normalised by its maximum value.

3.1. Binaural cues

The localization ambiguity arising from the cone of confusion can be resolved quite readily by head motion [1]. However, even if the head is restrained, partial resolution is still possible on the basis of the static spectral cues [19]. Resolution of the ambiguity is further improved if the listener has a priori information which restricts the possible source locations. For example, if the subject knows in advance that the sound source is in the horizontal plane in front. Considering these factors, it was necessary to verify the ability of the binaural unit of the model to resolve any static cues

of elevation, i.e. whether the EI-patterns are able to give any information of the location of an elevated source given that they only characterize the ITD and ILD cues.

Considering that the EI patterns depend on the frequency of the gammatone filterbank channel (k) of the peripheral processing unit, the azimuth ($\hat{\phi}$) and elevation angle ($\hat{\theta}$) of a target sound source, and the ITD (τ) and the ILD (α) that occurs due to the comparison of the signals of the left and the right ear, we compared the EI patterns created by a given $\hat{\phi}$ and $\hat{\theta}$ with all the EI patterns for all possible ϕ and θ in 3D space by using eq. 5.

In Figure 3 there are some representative results of the comparison of the EI patterns created by a white noise sound source at a given location $\hat{\phi}$, $\hat{\theta}$ in the vertical-polar coordinate system (fig. 2). From visual observation we can see that at low frequencies a clear circle is formed, which indicates a cone of confusion, and as a consequence, the inability of EI patterns to predict the location of elevated sources. Similar results have been obtained for frequencies up to 4kHz. This indicates that in low and middle range frequencies where the ITD cues are prominent, the EI patterns are not able to predict the location of elevated sources, however they give a clear indication of the lateralized sources. At higher frequencies as in figure 4 the circle is deformed. This indicates that at middle high frequencies where the ILD cues are more prominent, the EI patterns indicate a dependency on the elevated sources which could be explained by the fact that short-wavelength sounds are not diffracted around the head to the same extent as long wavelengths.

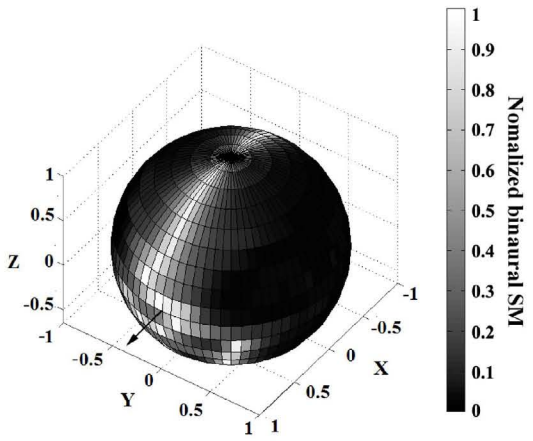


Figure 4: The results of the comparison of the EI patterns at $f \approx 9.4$ kHz for a sound source at $\hat{\phi} = \hat{\theta} = 0^\circ$ by using the vertical-polar coordinate system. The colour bar indicates the value of eq. 5 normalised by its maximum value.

3.2. Monaural cues

One of the main characteristics in the analysis of the head related transfer functions (HRTFs) is the spectral colouration introduced by the outer ear. Prominent peaks and notches can be found at different frequency ranges that are considered potential cues for elevation. For instance, the ambiguity on a cone of confusion can be discriminated with the appropriate spectral cues that reside mainly at 8 - 16 kHz [20], while for up-down location the appropriate spectral cues reside mainly at 6 - 12 kHz [20].

Additionally, it has been shown that the tonotopic organization in the cochlea is preserved in the higher regions of the auditory path such as in the cochlea nucleus (CN) [21]. As a consequence, it was considered necessary to check whether the peaks and notches of the HRTFs could be preserved in the TA patterns (eq. 4).

In Figures 5, 6 we can see[1] from visual observations that all the pinna resonances and pinna nulls of the HRTFs are preserved in the TA patterns but in a rather smoothed out representation. This smooth representation of the TA patterns is due to the lower frequency resolution of the channels of the gammatone filterbank (100 frequency channels) compared to the finest resolution of the HRTFs and the compressive character of the square root compressor in the peripheral processing unit which changes the dynamic range of the signal.

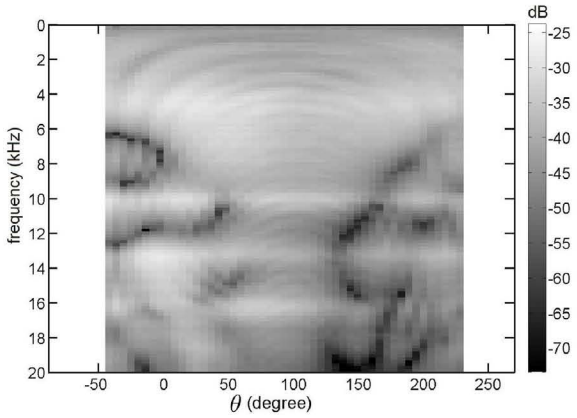


Figure 5: The HRTF of a KEMAR with a small pinna from the CIPIC database (subject 165, right ear) [16] in the median plane ($\phi = 0^\circ$) in the interaural-polar coordinate system.

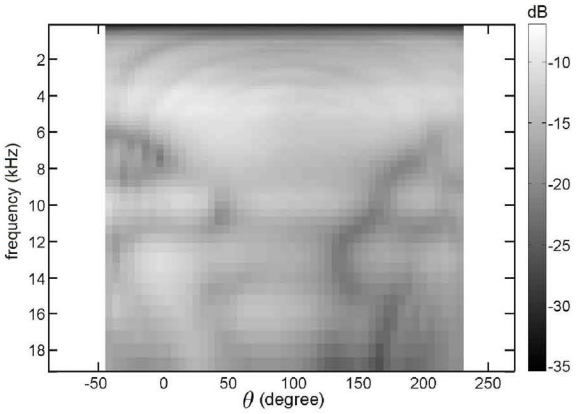


Figure 6: The TA patterns as they have been created by the HRTF of a KEMAR with a small pinna from the CIPIC database (subject 165, right ear) [16] in the median plane ($\phi = 0^\circ$) in the interaural-polar coordinate system.

[1] Although the results depict the monaural processor produced by the right ear, similar results could also be found at the corresponding TA patterns of the left ear.

3.3. Decision making device

Considering that the SM of the monaural and binaural cues have been combined as indicated in eq. 11 it was considered necessary to further explore the influence of the γ_{bin} (eq. 6) and σ_{mon} (eq. 9) parameters in the final stage of the model independently. Figures 7 - 10 illustrate the effect of the γ_{bin} and σ_{mon} in the binaural and monaural SMs for very high and very low values at a position exactly in front of a KEMAR ($\hat{\phi} = \hat{\theta} = 0°$), for a sound source as described in sec. 4. For the binaural SM, eq. 6, (Figures 7, 8), which is responsible for giving the highest similarity to all the points around the target azimuth angle independent of the elevation angle, it can be noticed that the γ_{bin} parameter spreads the values around the target azimuth angle $\hat{\phi} = 0°$. This implies that the binaural SM is roughly independent of the elevation angle ($s_{bin}(\phi, \theta) \approx s_{bin}(\phi)$) which indicates the lack of EI cues to match to all the EI patterns along the median plane.

In contrast, the monaural SM shows a different behavior. For high values of σ_{mon} (Figure 9), the TA cues around the median plane match with all the TA patterns indicating in this way a high chance the sound source is located at a position outside that region. Nevertheless at all locations the SM has a rather low value which ranges from 0.85-1.0. In cases where σ_{mon} is less than one (Figure 10) the performance of the monaural processor improves, and for extremely low values, the monaural processor gives the highest similarity at the point where a sound source is located.

Based on the behavior of the γ_{bin} parameter and the fact that the EI patterns are associated with the ITD and ILD cues, we could conclude that the binaural SM ($s_{bin}(\phi, \theta)$) is able to give an estimation of the position of the sagittal plane with the γ_{bin} parameter restricting or expanding the predicted region around the estimated sagittal plane. In addition, considering that TA patterns are associated with the spectral cues, the monaural SM is able to predict the exact location of a sound source with the σ_{mon} parameter restricting or expanding the predicted region around the estimated location. However, this is not only limited to the target position but it expands to other locations as well, where the TA patterns are quite similar. This is associated with the lack of the spectral cues to resolve the exact location of a sound source on a cone of confusion as indicated in Figure 10 where there is a high probability for a sound source at $\hat{\theta} = 180°$.

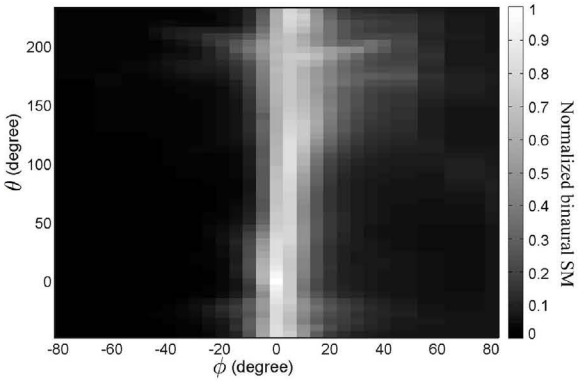


Figure 7: The prediction of the binaural pattern matching process (eq. 7) normalized by its maximum value for a white noise sound source as described in [22] at $\hat{\phi} = 0°$ and $\hat{\theta} = 0°$ in the interaural-polar coordinate system and for a high value of the γ_{bin} parameter ($\gamma_{bin} \gg 1$).

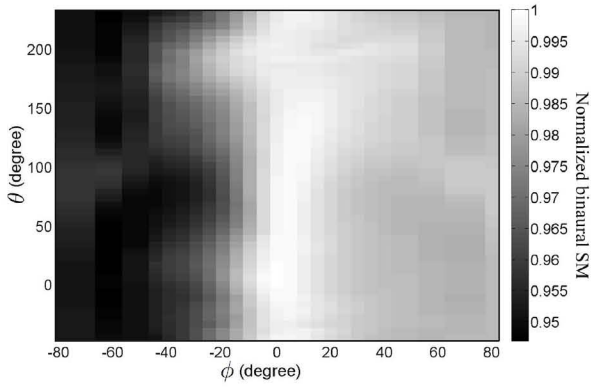


Figure 8: The prediction of the binaural pattern matching process (eq. 7) normalized by its maximum value for a white noise sound source as described in [22] at $\hat{\phi} = 0°$ and $\hat{\theta} = 0°$ in the interaural-polar coordinate system and for a low value of the γ_{bin} parameter ($\gamma_{bin} \ll 1$).

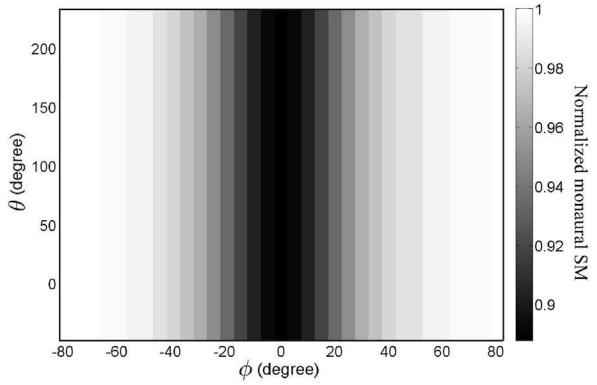


Figure 9: The prediction of the monaural pattern matching process (eq. 10) normalized by its maximum value for a white noise as a sound source at $\hat{\phi} = 0°$ and $\hat{\theta} = 0°$ in the interaural-polar coordinate system and for a high value of the σ_{mon} parameter ($\sigma_{mon} \gg 1$).

4. COMPARISON TO LISTENING TESTS

In order to validate the performance of the proposed model, the experimental data of Makous and Middlebrooks [22] have been used. In the particular listening test six listeners with normal hearing had to identify the actual location of a sound source at different locations in 3D space at a fixed distance of 1.2m in an acoustic environment with 40 dB SPL ambient noise and a room that can be considered anechoic for frequencies above 500 Hz. The sound source had a sound pressure level that ranged randomly for each trial from 40 to 50 dB sensation level and a frequency range between 1.8 kHz and 16 kHz. From the two experiments that were conducted we are mainly interested in the so called open-loop trials, in which the duration of the stimulus was 150ms and the subject had his/her head at a fixed position. In this way any dynamic cues that could have been created were excluded. Finally across all subjects, each stimulus location was tested in total 31 times giving an azimuth and elevation mean error and standard deviation for each subject.

Figures 11 - 13 illustrate the prediction of the model for a vir-

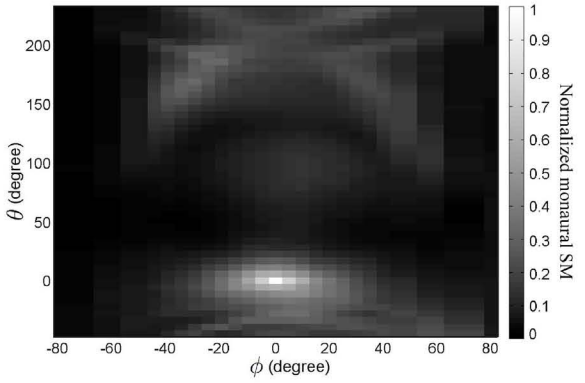


Figure 10: The prediction of the monaural pattern matching process (eq. 10) normalized by its maximum value for a white noise as a sound source at $\hat{\phi} = 0^\circ$ and $\hat{\theta} = 0^\circ$ in the interaural-polar coordinate system and for a low value of the σ_{mon} parameter ($\sigma_{mon} \ll 1$).

tual sound source[2] with the same specifications of the listening test[3] at three different positions. The center of the ellipses on the Figures indicate the average error of the detected sound source position in the listening tests and it has been calculated by averaging the mean error of the response of each subject. The average error is characterized by its mean value, which is the center of the ellipses, and the standard deviation about the mean value, which is not indicated. The length of transverse and conjugate diameters indicate the average standard deviation about the mean response for each subject for the azimuth and elevation angle correspondingly and it has been calculated by averaging the standard deviation of the response of each subject. The average standard deviation of the azimuth and the elevation angle is characterized by it's mean value, which is the length of transverse and conjugate diameters correspondingly, and a standard deviation about this value, which is not indicated. The parameters γ_{bin} and σ_{mon} of the model have been adjusted in such a way to fit as closely as possible to the listening test results, where $\gamma_{bin} = 1.82$ and $\sigma_{mon} = 0.3$.

Although the performance of the model, from visual observation of the figures 11 - 13, seem to give quite a good prediction of the results of the listening tests, some other aspects should be considered. Due to the fact that the frequency range of the sound source is between 1.8 kHz and 16 kHz all the information that is hidden in the low frequencies for the ITDs has been eliminated. This results in the total SM being spread along the estimated sagittal plane, something that is influenced by the fact that the EI patterns are only using the ILDs and the envelope of the ITD cues.

Despite the fact that the average error and the average standard deviation of the detected sound source position have been used for the creation of the ellipses of the listening tests, the actual errors are even higher. For instance for a sound source in the median plane at an elevated position at $\hat{\theta} = 45^\circ$ (Figure 11), the average error can vary from $2.7^\circ \pm 4.1^\circ$ for the horizontal dimension[4] and $-5.9^\circ \pm 10.6^\circ$ for the vertical dimension while the average standard deviation can vary from $3.0^\circ \pm 2.3^\circ$ for the horizontal dimension and $7.9^\circ \pm 2.0^\circ$ for the vertical dimension. This means that in general the error of the estimated location of the horizontal dimension can vary from -2.1° to 12.1° and from 18.6° to 59.6° for the vertical dimension. Furthermore, these estimated values do not consider the front-back confusion errors, something that is depicted by the prediction of the model.

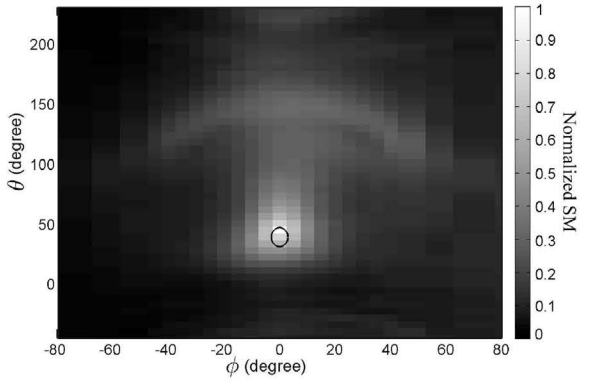


Figure 11: The prediction of the perceptual model (eq. 11) normalized by its maximum value for a sound source at $\hat{\phi} = 0^\circ$ and $\hat{\theta} = 45^\circ$ in the interaural-polar coordinate system and the listening test results (ellipse) of Makous and Middlebrooks [22].

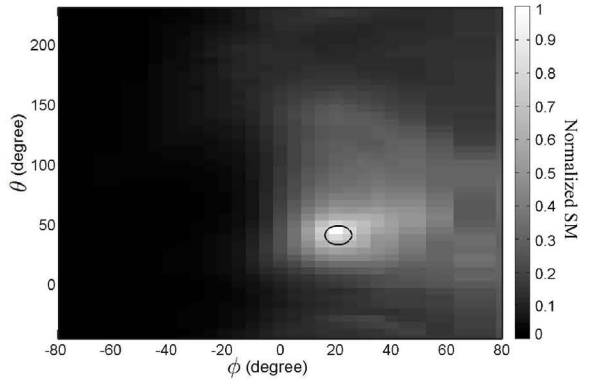


Figure 12: The prediction of the perceptual model (eq. 11) normalized by its maximum value for a sound source at $\hat{\phi} = 20^\circ$ and $\hat{\theta} = 45^\circ$ in the interaural-polar coordinate system and the listening test results (ellipse) of Makous and Middlebrooks [22].

5. CONCLUSIONS

The aim of the current study was to explore some of the characteristics of a biologically inspired model and to illustrate its performance in comparison to real listening tests. The results of the listening test indicate that the current model is able to predict, at least qualitatively, the human performance in localization tests of stationary sounds. Nevertheless, further investigation is necessary for a quantitative analysis of the model and a better quantification of the range that γ_{bin} and σ_{mon} should vary to predict the human performance in the localization of broadband sound sources with individualized or generalized HRTFs.

[2]The HRTFs that have been used are from the CIPIC database[16].

[3]The sound pressure level has been considered to be on average 45 dB SPL.

[4]In the notation $m \pm \sigma$ the first value indicates the mean value, while the second the standard deviation around this value.

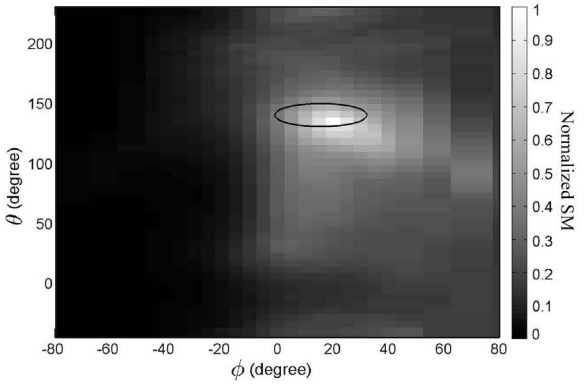


Figure 13: The prediction of the perceptual model (eq. 11) normalized by its maximum value for a sound source at $\hat{\phi} = 20°$ and $\hat{\theta} = 135°$ in the interaural-polar coordinate system and the listening test results (ellipse) of Makous and Middlebrooks [22].

6. ACKNOWLEDGMENTS

The research for this paper was financially supported by Meridian Audio Ltd. and the University of Southampton. In developing the ideas presented here, I have received helpful input from Dr. Stephan Bleeck from the University of Southampton, and Prof. Ville Pullki from the Aalto Department of Signal Processing and Acoustics. Very many thanks also to Dr. T. Takeuchi, Dr. M. Park and Prof. J. C. Middlebrooks , amongst others, for useful feedback and advice.

7. REFERENCES

[1] J Blauert, *Spatial Hearing: The Psychophysics of Human Sound Localization*, MIT, Cambridge, MA, 1997.

[2] Lloyd A Jeffress, "A place theory of sound localization," *Journal of Comparative and Physiological Psychology*, vol. 41, no. 1, pp. 35–39, 1948.

[3] N. I. Durlach, "Equalization and Cancellation Theory of Binaural Masking-Level Differences," *The Journal of the Acoustical Society of America*, vol. 35, no. 8, pp. 1206–1218, 1963.

[4] G. Von Bekesy, "Zur Theorie des Hörens: Über das Richtungshören bei einer Zeitdifferenz oder Lautstärkeungleichheit der beidseitigen Schalleinwirkungen," *Phys Z*, pp. 824–838, 1930.

[5] Erno H A Langendijk and Adelbert W Bronkhorst, "Contribution of spectral cues to human sound localization," *The Journal of the Acoustical Society of America*, vol. 112, no. 4, pp. 1583–1596, 2002.

[6] Robert Baumgartner, Piotr Majdak, and Bernhard Laback, "Assessment of sagittal-plane sound-localization performance in spatial-audio applications," in *The technology of binaural listening*, pp. 93–120. 2013.

[7] Paul M Hofman and A John Van Opstal, *Spectro-temporal factors in two-dimensional human sound localization*, vol. 103, ASA, 1998.

[8] Joyce Vliegen and A John Van Opstal, "The influence of duration and level on human sound localization.," *Journal of the Acoustical Society of America*, vol. 115, no. 4, pp. 1705–1713, 2004.

[9] Munhum Park, Philip A. Neslon, and Kyeongok Kang, "A Model of Sound Localisation Applied to the Evaluation of Systems for Stereophony," *Acta Acustic United with Acustica*, vol. 94, pp. 825–839, 2008.

[10] R.D. Patterson, K. Robinson, J. Holdsworth, D. McKeown, C. Zhang, and M. Allerhand, "Complex sounds and auditory images," in *Auditory physiology and perception, 9th International Symposium on Hearing*, Y. Cazals, L. Demany, and K. Horner, Eds., Oxford, 1992, pp. 429–446, Pergamon, 1992.

[11] Donald D Greenwood, "What is "Synchrony suppression"?," *The Journal of the Acoustical Society of America*, vol. 79, no. 6, pp. 1857–1872, 1986.

[12] R C Kidd and T F Weiss, "Mechanisms that degrade timing information in the cochlea.," *Hearing Research*, vol. 49, no. 1-3, pp. 181–207, 1990.

[13] T Dau, D Püschel, and A Kohlrausch, "A quantitative model of the "effective" signal processing in the auditory system. I. Model structure.," *Journal of the Acoustical Society of America*, vol. 99, no. 6, pp. 3615–3622, 1996.

[14] Sergios Theodoridis and Konstantinos Koutroumbas, *Pattern Recognition*, Academic Press, 4th edition, 2009.

[15] Symeon Mattes, Philip Arthur Nelson, Filippo Maria Fazi, and Michael Capp, "Towards a human perceptual model for 3D sound localization," in *28th Conference on Reproduced Sound: Auralisation: Designing With Sound*, 2012.

[16] V. R. Algazi, R. O. Duda, D. M. Thompson, and C. Avendano, "The CIPIC HRTF database," in *Proc. 2001 IEEE Workshop on Applications of Signal Processing to Audio and Electroacoustics*, pp. 99–102, 2001.

[17] Richard M Stern, Andrew S Zeiberg, and Constantine Trahiotis, "Lateralization of complex binaural stimuli: A weighted-image model," *The Journal of the Acoustical Society of America*, vol. 84, no. 1, pp. 156–165, 1988.

[18] Ewan A Macpherson and Andrew T Sabin, "Binaural weighting of monaural spectral cues for sound localization.," *Journal of the Acoustical Society of America*, vol. 121, no. 6, pp. 3677–3688, 2007.

[19] Frederic L Wightman and Doris J Kistler, "Monaural sound localization revisited," *The Journal of the Acoustical Society of America*, vol. 101, no. 2, pp. 1050–1063, 1997.

[20] Henrik Møller and Daniela Toledo, "The Role of Spectral Features in Sound Localization," *Audio Engineering Society Convention 124/7450*, 2008.

[21] R E Wickesberg and D Oertel, "Tonotopic projection from the dorsal to the anteroventral cochlear nucleus of mice.," *Journal of Comparative Neurology*, vol. 268, no. 3, pp. 389–399, 1988.

[22] James C Makous and John C Middlebrooks, "Two-dimensional sound localization by human listeners," *The Journal of the Acoustical Society of America*, vol. 87, no. 5, pp. 2188–2200, 1990.

A ROADMAP FOR ASSESSING THE QUALITY OF EXPERIENCE OF 3D AUDIO BINAURAL RENDERING

Rozenn Nicol, Laetitia Gros, Cathy Colomes
Orange Labs
Lannion, France
rozenn.nicol@orange.com

Olivier Warusfel, Markus Noisternig, Hélène Bahu
UMR STMS IRCAM-CNRS-UPMC
Paris, France
olivier.warusfel@ircam.fr

Brian FG Katz, Laurent S. R. Simon
LIMSI-CNRS
Orsay, France
brian.katz@limsi.fr

ABSTRACT

Today there are 2 major evolutions in spatial audio. First, an enhanced 3D audio experience, where virtual sound sources can be accurately synthesized in any direction, is possible with technologies such as binaural, Wave Field Synthesis, Higher Order Ambisonics or Vector Base Amplitude Panning. Second, 3D audio is on the way to being democratized through binaural adaptation for headphone listening. These evolutions call for revisiting the methods and tools used to assess the perception of spatial sound reproduction. The first objective of this paper is to delineate the problem, by exploring the potential dimensions and the related attributes underlying the perception of spatial sound, mainly within the context of binaural reproduction. Secondly, assessment methods, including both standard and less conventional ones, are listed, and their relevance for the measure of the attributes previously identified is discussed.

1. INTRODUCTION

Sound spatialization is undergoing major evolutions with promises of enhanced 3D audio experience and the recent inclusion of height information. Beyond discrete channel audio, sound field representation formats such as Wave Field Synthesis (WFS) and Higher-Order Ambisonics (HOA) are now appearing in the professional community. These technologies typically require large sophisticated installations. However, binaural audio playback using Head Related Transfer Functions (HRTFs) makes 3D audio immersion possible for any listener using nothing more than standard headphones.

The use of 3D audio technologies allows for synthesizing virtual sound sources at any position in space with an accuracy that is very close to natural listening [1]. Binaural audio with headphones could be the first technology which makes immersive 3D audio mass consumable. Spatial information becomes then an inherent feature of the experience. This raises the question of to what extent we can quantify and optimize the quality of experience (QoE) in immersive audio systems.

The consequence is that methods and tools for audio assessment should be revisited. At least, a deeper investigation of spatial attributes is required. Perceptual attributes, such as "spatial impression", "depth", "envelopment", "width" [2, 3], which were identified by past studies on room acoustics and linked to spatial perception, are no longer sufficient. More generally, our main concern here is to measure how a 3D sound scene is perceived by a listener, whatever technologies or listening setups are used. The objective is to assess the perceived "quality", i.e. to identify the perceptual dimensions used by the listeners to make their judgments. Two main categories of dimensions are already identified: 1- the spectral content (i.e. timbre), 2- the spatial location of the sound [4]. More global audio quality ratings rely on other perceptual dimensions which have to be investigated, and for which new tools of assessment are yet to be identified. Naturalness, or plausibility, of the virtual sound sources is one example of such additional dimensions.

It is beyond the scope of this article to answer these questions on a global scale. In the following we will mainly focus on the context of binaural sound reproduction, in the case of a direct binaural synthesis of the different audio objects of a sound scene as well as in the case of the binaural decoding of any channel based audio format using the virtual speaker paradigm. It is well known from literature, for instance, that the use of non-individualized HRTFs degrades both the spectral and spatial "quality" of the reproduced sound sources. Besides the use of individual or generic HRTFs, the decoding of a stereo or surround format through the virtual speaker paradigm is also less convincing than a direct binaural rendering of the sound sources. Thus our concern may be the assessment of the "quality" of any given set of HRTFs. By "quality" it is meant to measure how a source is perceived by a listener when processed by a given HRTF or how a sound scene is perceived when rendered or decoded through binaural synthesis. Usually the quality of HRTFs is assessed by localization tests, which mainly focus on the localization accuracy and often do not measure other perceptual dimensions (such as timbre).

Another conventional method of assessment is the measure of Basic Audio Quality (BAQ) [5], where "Quality" here refers to the fidelity with which a signal is transmitted or rendered by a system. In other words, degradations are measured in comparison to a given reference. It has been widely used in perceptual audio experiments as a quick means to compare different alterations of a signal, and was initially developed to classify audio codecs. Applying them to the assessment of the quality of HRTFs is questionable. Assessing binaural reproduction and in particular the suit-

ability of a set of HRTFs can be considered as a measure of degradation. In most cases, the reference is unknown to the listener, unless the subject can directly compare the real audio scene (e.g., sound played from a real loudspeaker at the same position as the virtual sound source, played from a virtual loudspeaker) with its binaural representation. In addition, evaluating BAQ gives little, if any, information about the artifacts of a given binaural technique. The exception to this rule is the assessment of modeled HRTFs, in which the reference is the measured individual HRTF. However, this choice of reference is not necessarily the ideal case, insofar as real world HRTF measurements are not free from errors and approximations. Besides, the claim that measured individual HRTFs provide the best overall audio experience (i.e. not only in terms of localization, but for all the other perceptual dimensions) remains to be verified.

We propose using the term Quality of Experience (QoE) as an alternative to BAQ. QoE measures how a subject experiences a given system. Since the range of situations that can be covered by binaural reproduction is very large, we believe this QoE should remain multidimensional and multicontextual.

The remainder of this article is an attempt to draw a roadmap of the different dimensions and methods that should be investigated to delineate the numerous dimensions of the Quality of Experience. In the following, two principal questions will be addressed. Section 1 will first explore what are the potential dimensions underlying the perception of binaural sound. Then, Section 2 will present an overview of available methods of perceptual assessment. Both conventional and new tools will be considered. Three different groups of methods to assess the perceptual audio quality will be used for classification: a) direct assessment of a perceptual attribute without any reference; b) direct assessment with a reference; and c) indirect assessment, in which case the quality is inferred from the subject's behavior (e.g., measure of task performance). For each perceptual dimension the relative merits of the different methods will be discussed.

2. PERCEPTUAL DIMENSIONS OF BINAURAL SOUND

Binaural audio consists of a left and right ear signal that can be directly played back over headphones. These signals can be directly recorded (e.g., using a dummy head or in-ear microphones), or synthesized using individual or non-individual HRTF filters. QoE addresses the questions of a) how can a listener describe his/her perception, and b) what are the objective features (especially acoustical) and how do they correspond to the perceptual dimensions. Some dimensions are already known. For instance, the physical properties of the sound scene that the binaural sound intends to reproduce have clearly an influence on perception, namely: the frequency content of the sound signal, the location of the sound source, the acoustic environment (room effect), etc. The perceptual attributes related to these physical parameters are called "physical-related attributes" in the following discussions. Another category of attributes concerns the effect on the psychic or affective state of the listener: are the virtual sound sources plausible, to what extent does the listener feel immersed in the virtual sound scene, what are his/her emotion(s), etc.? More generally, perceptual studies need to be reconsidered in order to take into account the specific context of listening, involving perception-action feedback according to a given task or cognitive situation [6].

2.1. How to investigate perceptual dimensions?

Different methods to identify the perceptual dimensions and associated attributes have been proposed in the literature. The two most prevalent approaches are: 1- Multi-Dimensional Scaling (MDS) and 2- verbal elicitation techniques, e.g. Descriptive Analysis (DA), Repertory Grid Technique (RGT), or Free-Choice Profiling (FCP). MDS measures the perceptual dissimilarities between a large set of stimuli and derives the most relevant perceptual parameters from these distances. However, MDS methods do not always guarantee that the dimensions revealed by the analysis correspond actually to perceptual attributes. The INDSCAL method has been introduced to overcome this limitation [7] and has been used for musical timbre analysis [8] as well as for perceptual studies in room acoustics [9]. Alternatively, DA, RGT and FCP are direct verbal methods for eliciting and evaluating people's subjective experiences and perceived differences between the stimuli of a similar large set of stimuli [10, 11]. They originate from the food industry, where they've been used for creating dictionaries of words for flavour qualification. These methods help in constructing a space corresponding to the perceptual attributes elicited in the given set of stimuli, and can also infer the words that best describe each end of a perceptual dimension. However, DA, RGT, and FCP are time-consuming methods, both for the experimentor as well as for the subjects of the experiment.

According to [12], the first step of dimensional analysis methods is to generate a large set of stimuli that is representative of the differences encountered in the area of interest, so that many or all perceptual differences can be expected to be found when comparing all the stimuli.

In the context of binaural sound reproduction, building a set of representative stimuli is a non-trivial task. Binaural stimuli are highly individual: HRTFs vary widely from one subject to another and using non-individual HRTFs leads to large perceptual differences. Sets of stimuli that include non-individual binaural signals will therefore be perceived differently from one subject to another, and may lead to different constructs from one subject to another, making the statistical analysis of the elicitation process more complex. A solution would be to generate a set of sufficiently different stimuli so that similar constructs are found. It would therefore need to include individual HRTFs, deteriorated individual HRTFs and non-individual HRTFs. The deteriorations would need to be perceptually identical from one subject to another, which is not possible as long as the perceptual attributes and their underlying models remain unknown. Finding the perceptual dimensions in the context of binaural recordings should therefore be considered as an iterative process, where several consecutive experiments of perceptual dimension identification should be conducted. In these conditions, an exhaustive identification of all the dimensions of the perception of binaural sound is a very difficult task, if possible at all.

2.2. Physical-related attributes

Physical-related attributes describe perceptual attributes that can be directly linked to a physical or mathematical property of either the sound source, the acoustic space, or the sound reproduction system.

2.2.1. Timbral attributes

According to the British Oxford dictionary, timbre is *the character or quality of a musical sound or voice as distinct from its pitch and intensity*. According to this definition, spatial characteristics should be part of a sound's timbre. Indeed the acoustic response of the room contributes to the timbre of sound at the listener's place. At different positions in a room listeners will not perceive the same timbre of a sound. What's more, in binaural reproduction, timbre has an ambiguous role: spectral features are partly interpreted as localization cues. However, a number of studies separate timbre properties of a sound from its spatial properties [13, 4]. In this paper, we will consider timbral attributes separately from spatial attributes, though some attributes may overlap.

Several lists of timbral attributes have been designed. [14] proposed a list of timbral attributes that can be used in machine learning. However, attributes such as the zero crossing rate or centroid temporal peakedness can hardly be used by subjects to describe timbral qualities of sound. For this reason, lists of specific perceptual attributes of timbre have been developed for music instruments [15], speech [16], and loudspeakers [10]. Timbral attributes are generally related to the spectro-temporal properties of the sound.

2.2.2. Source location

The location of a sound source is, to some extent, directly perceptible by listeners. It is generally expressed in terms of azimuth, elevation and distance. Spherical coordinate systems are preferred to cartesian coordinate systems as the sound source position is perceived relatively to the subject itself. Localizability, or spatial definition, refers to the ease of localizing the sound and is an additional aspect worth considering.

Perception of distance is rarely assessed in listening tests (stereophonic or multichannel reproduction). However, it should be highlighted that, in the specific case of binaural sound, distance perception is of greater importance. This is related on one hand to the ability of binaural reproduction to render varying distances and on the other hand to the problem of inside-the-head localization (IHL) and externalisation as a common artifact of binaural reproduction is that the virtual sound source is localized inside or close to the head. Therefore, when judging auditory distance for binaural sound, this phenomena must be carefully examined.

2.2.3. Perceived width and Apparent Source Width (ASW)

The perceived width of a sound source is the measure of an auditory event's spatial extent, which can be expressed in terms of an angular span (i.e. spherical sector) and depth. An auditory event does not usually have clear limits, which makes it difficult to define without a reference. In room acoustics, Apparent Source Width (ASW) is affected by from where and at what times the early reflections arrive [17, 18, 19]. ASW is also affected by the physical extent of the source (when not a point source) and to the sound reproduction system [20, 21]. Objective measures have found ASW to be correlated to the interchannel or interaural cross-correlation [22, 23]. Several other attributes are perceptually close to ASW yet different enough to require distinction: locatedness [24] and diffuseness of the source, as well as spatial unity

2.2.4. Room-related attributes

The room in which sounds are played and/or recorded has a strong influence on their perceptual attributes. Rooms add reflections to the original sound (the direct sound), which may result in a change in perceived timbre or spatial properties. The source directivity also influences the perceived sound [25]. Some spatial attributes and their characterization can be found in, e.g., [17, 26, 10]. However, the number of perceptual dimensions remains unclear. Early work of Sabine identified 3 attributes ("loudness", "distorsion of complex sounds: interference and resonance", and "confusion: reverberation, echo and extraneous sounds"). Present literature rely on at least 7-9 necessary attributes [27, 28]. The study of their correlation with acoustic criteria showed that the perception of room quality is mainly influenced by the energy of the direct sound (including early reflections), the overall energy of reverberated sound, the decay time of reverberation (i.e. reverberation time), the time and spatial distribution of early reflections, and the frequency balance of each criteria, all of which have well-defined metrics.

2.2.5. Discussion

Several physical parameters which potentially represent a sound scene are listed in the previous section. For some, it is unknown if and how they affect auditory perception. The intensity of the perceptual effect is probably not the same for each parameter: their relative weights need also to be determined. In addition, this physical description should be revisited in light of studies concerning auditory scene analysis [29]. The brain distorts the "physical reality" for building the associated percept (concept of auditory streams based on grouping or segregation of auditory events). Information is ordered by mental and cognitive processing, sometimes independently from the physical properties. Some studies suggest for instance that frequency features are of primary importance, above spatial properties [29].

2.3. Psychic and affective attributes

Other attributes may be related more to the listening experience than to physical properties of the sound sources or the properties of the room. Psychic and affective attributes refer to the results of further processing and analysis of the sound scene by the brain. That which is of interest is no longer the "pure acoustic information", but the way in which the psychic state of the listener is modified by the sound. This question is rather new, and we are far from having a clear understanding of all the dimensions involved. It should be noted however that these effects are obviously highly dependent on the audio content and personal experience of the subject. As a first contribution, we propose here to consider 3 potential attributes: naturalness (and its correlates), readability, and emotion.

2.3.1. Naturalness

Binaural technology offers the ability to reproduce at the ears of a listener the exact sound that would have arrived at his/her ears if he/she had been located in the original environment. Alternatively, binaural synthesis can be used to create an auditory environment, in which case the only reference a listener may have is an expectation of the auditory scene which might be induced from memory or derived from visual information. Since binaural technology aims at mimicking natural listening, the realism, the naturalness, and

the fidelity to the original sound scene or to the expected one, often seems to be a more important question in this context than for other spatial sound technologies.

[30] showed that naturalness is a desirable characteristic of an environment. In addition, one can expect to find naturalness ratings being strongly correlated to preference ratings, as [31] showed. Thus the rating of naturalness might be influenced by preference, even if a preference judgment is not explicitly asked from the listeners. Nevertheless, in some situations, naturalness might not be as desirable as listeners may think. For instance, it happens that, when attending a concert, our natural experience is poor (e.g., the sound scene is perceived as ill-defined and narrow), maybe due to the listening's position. A recording of such a sound scene could then be poorly rated in terms of preference and possibly in terms of realism, in contrast to a more processed recording, where each individual source is clearly definable, and which could be perceived as more natural even though it would be physically impossible.

Naturalness of a sound scene is regularly used in perceptual evaluations [10] but is should be treated carefully. Naturalness is essentially a comparison between an unknown reference (the original sound scene, where the listener often was not present or has an old and potentially erroneous memory of it) and a known signal (the binaural reproduction). [32] questions the desirability of fidelity and naturalness. Nuances of the concept of naturalness, such as plausibility and presence, have been developed to address additional perceptual attributes.

2.3.2. Readability

When listening to a complexe sound scene, i.e. composed of many sources conveying various information, the question of its readability is an important parameter. Here this term refers to the ability to discriminate the different concurrent sound sources, in order to focus on one specific component [33]. Speech intelligibility is one particular example (i.e. cocktail-party effect [34]), but readability it pertinent for any audio content. In classical music, it describes the ability of the listener to dynamically focus on one instrument (or group of instruments) [35]. More generally, readability is one aspect of auditory scene analysis, which allows one to separate the overall scene into several streams with various levels of processing. Readability is affected for instance by frequency or spatial separation.

2.3.3. Emotion

The emotional dimension corresponds to any emotion that is felt by the listener, whether positive or negative. Our primary concern is to acknowledge that the listener is "touched" by the sound, which is an indication of a certain degree of immersion. Nevertheless, the nature of the emotion is also relevant information. After Wundt [36], emotion is described by 3 dimensions: valence (pleasure vs unpleasure), vigilance or dominance (no control vs maximal control), and arousal (excitation vs relaxation). However, valence and arousal have been shown as the more reliable dimensions [37]. Measuring and analyzing emotions is a difficult task, all the more that it is highly dependent on the content and the individual, though some bio-metric data has been shown to provide correlations to certain aspects.

3. WHICH EXPERIMENTAL PROCEDURE FOR WHICH PERCEPTUAL PARADIGM?

Evaluating the above-mentioned perceptual attributes requires suitable experimental protocols. This section and the following one review the most common approaches to perceptual evaluation in the context of binaural experience and discuss for each one the perceptual paradigms they may help to evaluate. From experimental psychology, two main categories of evaluation are distinguished: direct evaluation, where the subject is directly asked to rate the attributes under study, and indirect evaluation, where the subject's perceptual rating is inferred [12].

3.1. Direct measure of attributes

This method intends to assess each perceptual attribute separately. For example, if we consider the perception of timbre and focus on the descriptor "clarity" [38], the subject is asked to rate the clarity of audio stimuli. For each stimulus, he / she gives a score within an appropriate scale ranging between 0%, meaning a muffled sound, and 100%, meaning a clear sound. For each attribute, the choice of the scale (grading, labeling) must be carefully designed as a function of the considered attribute. A single attribute or a set of attributes can be measured at the same time. Another example is localization accuracy testing, where the subject is asked to report the azimuth, elevation, and/or distance of a sound source. In that case, the scales are in degrees or meters. A third example is the assessment of room quality by a questionnaire, in which the subject is asked to rate a set of attributes ("presence", "room effect", "reverberance", etc.).

3.2. Direct measure of attributes with a reference

A reference is defined here as any audio or visual stimulus which is provided to the subject and to which he/she is asked to compare to the signal under assessment. It is not necessarily a reference representative of a high quality standard. The task of the subject is to rate an attribute of an audio sample in comparison to the given reference. In addition, specific anchors, corresponding to various grades along the judgment scale, can be included. This inclusion allows for a check of the reliability of the subject (median anchor) or to insure that the listeners are using the scale consistently through the whole experiment (low anchor). Anchors should be chosen so that the stimuli are distributed along the whole scale that listeners have to use. If not, one may encounter various kinds of biases as described in [39]. [40] describes two types of anchors: direct anchors, that are explicitly given to the listener, similar to references, and indirect or "hidden" anchors, which listeners are not aware of. The main advantage of direct anchoring is that it can help stabilize the range of estimations given by the subjects. Indirect anchors can also be used for that, though less efficiently. Indirect anchors are useful to estimate the reliability of the subject or biases related to the non-uniformity of the results.

Examples of direct measure with a reference are AB or ABX tests, recommended when the differences between stimuli are small, or MUSHRA tests [41] if the differences are moderate, though they were designed to evaluate degradations caused by audio codec compression. A third example is a relative localization test through alignment, where the listener needs to match the perceived direction or distance of a test stimulus with that of a reference stimulus.

3.3. Indirect measure

3.3.1. Task performance

In this situation, the subject is asked to perform a task in the context of binaural sound. His/her QoE is inferred from his/her success. For instance, the listener's task can consist in describing the sound scene, that is to report the number, the nature, and the location of the sound sources [42]. Another example is given by [43] who inferred QoE by asking the listener to explore the sound scene and find targets. In [44], Guillon described a localization test where the listener had to localize the virtual sound source as fast as possible, recording the response time for various sets of HRTFs corresponding to different levels of HRTF modeling. Results showed a correlation between the response time and the modeling quality. The general intent of these types of experiments is to derive information about the naturalness and readability of the sound scene from observations of the listener's behavior.

3.3.2. Physiological measures

Psychophysiology studies the relationships between physiological responses and psychological changes. The principle is to observe cognitive, emotional, or behavioral phenomena by analyzing the physiological responses of the subject. Electrodermal activity, heart pulse, skin temperature, or eye activity are examples of physiological observables that can be recorded and linked to the psychic state of the listener [45, 46]. Particularly, electrodermal activity and heart pulse are considered as relevant measure of emotions, the former being associated to the "arousal" dimension, the latter to the "valence" dimension.

3.3.3. Brain imagery

Magnetic Resonance Imagery (MRI), electroencephalogram or magnetoencephalogram are useful tools for observing brain activity. Particularly, technological progress has made electroencephalograms easier to measure with a simple headset. Understanding the link between neuronal activity and the underlying mental process is improving daily. Currently, we are not able to translate brain activity maps directly into what subjects think or feel. However, some information about their emotions can be inferred from knowledge of neuronal activity and connections. For instance, in [47], brain activity of ferrets listening to virtual sounds was compared between individual and non-individual HRTFs, showing that the spatial selectivity of neurons is strongly altered in the case of non-individual synthesis. This is a potential measure of naturalness or plausibility. Brain activity can also give information on the mental effort required by the listening task, which could be linked to readability. This measure could be of interest when subjects are sound engineers being observed during a post-production session. Brain imagery appears thus as a promising tool to investigate perception of spatial sound in general, and binaural sound in particular.

4. CONCLUSIONS

Assessment of the QoE of spatial sound reproduction is the question raised by this article. Rather than providing answers, this paper attempts to clarify what is known and what requires further investigations, for which several promising tools are recommended. Complementary assessments are needed: on the one hand, overall ratings where all dimensions are taken into account, and on the other hand, unidimensional measures focused on one specific attribute. In order to accomplish this however, the first step is to clarify the number and semantic interpretation of the numerous perceptual dimensions. An open question is whether these dimensions depend on the sound reproduction system, namely the audio spatialization technology used. In addition, it is important to identify or develop objective criteria correlated with the perceptual dimensions.

5. ACKNOWLEDGMENTS

This work was carried out within the context the FUI funded project "BiLi" (www.bili-project.org) with support from "CAP DIGITAL - PARIS REGION".

6. REFERENCES

[1] McAnally K. Martin R. and Senova M., "Free-field equivalent localization of virtual audio," *J. Audio Eng. Soc.*, vol. 49, pp. 14–22, 2001.

[2] "ITU-R BS.1284-1: general methods for the subjective assessment of sound quality," Tech. Rep., 2003.

[3] "EBU Tech 3286 assessment methods for the subjective evaluation of the quality of sound programme material - music," Tech. Rep., 1997.

[4] Sarah Le Bagousse, Mathieu Paquier, Catherine Colomes, and Samuel Moulin, "Sound quality evaluation based on attributes - application to binaural contents," Oct. 2011, Audio Engineering Society.

[5] "ITU-R BS.1387-1: method for objective measurements of perceived audio quality," Tech. Rep., 2001.

[6] J. Blauert, D. Kolossa, K. Obermayer, and K. Adiloğlu, "Further challenges and the road ahead," in *The Technology of Binaural Listening*, Jens Blauert, Ed., Modern Acoustics and Signal Processing, pp. 477–501. Springer Berlin Heidelberg, Jan. 2013.

[7] J. D. Carroll and J. J. Chang, "Analysis of individual differences in multidimensional scaling via an n-way generalization of "eckart-young" decomposition," *Psychometrika*, vol. 35, pp. 283–319, 1970.

[8] J.M. Grey, "Multidimensional perceptual scaling of musical timbres," *J. Acoust. Soc. Am.*, vol. 61, pp. 1270, 1977.

[9] C. Lavandier, *Validation perceptive d'un modèle objectif de caracérisation de la qualité acoustique des salles*, Ph.D. thesis, Université du Maine, Le Mans, 1989.

[10] Nick Zacharov and Kalle Koivuniemi, "Unravelling the perception of spatial sound reproduction: Language development, verbal protocol analysis and listener training," Nov. 2001, Audio Engineering Society.

[11] Gaetan Lorho, *Perceived quality evaluation: an application to sound reproduction over headphones*, Ph.D. thesis, Aalto University, Aalto, 2010.

[12] Nick Zacharov and Soren Bech, *The Perceptual Audio Evaluation: Theory, Method and Application*, John Wiley & Sons, 2006.

[13] Francis Rumsey, Slawomir Zielinski, Rafael Kassier, and Soren Bech, "On the relative importance of spatial and timbral fidelities in judgments of degraded multichannel audio quality," *Journal of the Acoustical Society of America*, vol. 118, no. 2, pp. 968–976, Aug. 2005.

[14] James McDermott, Niall J. L. Griffith, and Michael O'Neill, "Timbral, perceptual, and statistical attributes for synthesized sound," *IN: PROC. OF THE INTERNATIONAL COMPUTER MUSIC CONFERENCE. (2006*, 2006.

[15] Taffeta M. Elliott, Liberty S. Hamilton, and Frédéric E. Theunissen, "Acoustic structure of the five perceptual dimensions of timbre in orchestral instrument tones," *The Journal of the Acoustical Society of America*, vol. 133, no. 1, pp. 389–404, Jan. 2013.

[16] Ville-Veikko Mattila, "Descriptive analysis of speech quality in mobile communications: Descriptive language development and external preference mapping," Nov. 2001, Audio Engineering Society.

[17] David Griesinger, "General overview of spatial impression, envelopment, localization, and externalization," in *15th Audio Engineering Society International Conference: Audio, Acoustics & Small Spaces*, Copenhage, Denmark, Oct. 1998, Preprint 15-013.

[18] David Griesinger, "Objective measures of spaciousness and envelopment," in *Audio Engineering Society 16th International Conference: Spatial Sound Reproduction*, Rovaniemi, Finland, Mar. 1999, Preprint 16-003.

[19] Ingo B. Witew and Johannes A. Buechler, "The perception of apparent source width and its dependence on frequency and loudness," *The Journal of the Acoustical Society of America*, vol. 120, no. 5, pp. 3224–3224, Nov. 2006.

[20] B F G Katz and C. d'Alessandro, "Apparent source width and the church organ," in *7ème congrès de la Société Française d'Acoustique et 30ème congrès de la Société Allemande d'Acoustique (CFA/DAGA 2004)*, Strasbourg, 2004, pp. 1235–1236.

[21] Kimio Hamasaki, Toshiyuki Nishiguchi, Hiroyuki Okubo, Yasushige Nakayama, Reiko Okumura, and Masakazu Iwaki, "Natural reproduction of symphony orchestra music by an advanced multichannel live sound system," in *121st Audio Engineering Society Convention*, Oct. 2006, Preprint 6966.

[22] Russell Mason, Tim Brookes, and Francis Rumsey, "Development of the interaural cross-correlation coefficient into a more complete auditory width prediction model," in *Proceedings of the 18th International Congress on Acoustics*, Kyoto, Japan, 2004, vol. IV, pp. 2453–2456.

[23] Toshiyuki Okano, Leo L. Beranek, and Takayuki Hidaka, "Relations among interaural cross-correlation coefficient (IACCE), lateral fraction (LFE), and apparent source width (ASW) in concert halls," *The Journal of the Acoustical Society of America*, vol. 104, no. 1, pp. 255–265, July 1998.

[24] Thomas Lund, "Enhanced localization in 5.1 production," in *Audio Engineering Society 109th convention*, Los Angeles, United States, 2000, Preprint 5243.

[25] Markus Noisternig, Franz Zotter, and Brian FG Katz, "Reconstructing sound source directivity in virtual acoustic environments," in *Principles and Applications of Spatial Hearing*, Yôiti Suzuki, Douglas S Brungart, and Hiroaki Kato, Eds., pp. 357–373. World Scientific Publishing Co. Pte. Ltd., 2011.

[26] Jan Berg and Francis Rumsey, "Identification of perceived spatial attributes of recordings by repertory grid technique and other methods," in *Audio Engineering Society 106th convention*, Münich, May 1999, Preprint 4924.

[27] L. L. Beranek, "Concert hall acoustics," *J. Acoust. Soc. Am.*, vol. 92, pp. 1–39, 1992.

[28] E. Kahle, *Validation d'un modèle objectif de la perception de la qualité acoustique sur un ensemble de salles de concert et d'opéra*, Ph.D. thesis, Université du Maine, Le Mans, 1995.

[29] A. S. Bregman, *Auditory scene analysis: The perceptual organization of sound*, The MIT Press, 1990.

[30] Paul Rozin, "The meaning of "Natural". process more important than content," *Psychological Science*, vol. 16, no. 8, pp. 652–658, 2005.

[31] Jan Berg and Francis Rumsey, "Correlation between emotive, descriptive and naturalness attributes in subjective data relating to spatial sound reproduction," in *109th Audio Engineering Society Convention*, Sept. 2000, Preprint 5206.

[32] Francis Rumsey, "Faithful to his master's voice? questions of fidelity and infidelity in music recording," in *Recorded Music: Philosophical and Critical Reflections*. 2008.

[33] Catherine Guastavino and Brian F G Katz, "Perceptual evaluation of multi-dimensional spatial audio reproduction," *The Journal of the Acoustical Society of America*, vol. 116, no. 2, pp. 1105–1115, Aug. 2004, PMID: 15376676.

[34] A. W. Bronkhorst, "The cocktail party phenomenon: A review of research on speech intelligibility in multiple-talker conditions.," *ACUSTICA - acta acustica*, vol. 86, pp. 117–128, 200.

[35] J. Sanson, *Contrôle musical et perceptif de la spatialisation sonore en zone étendue*, Ph.D. thesis, UPMC, Paris, 2011.

[36] W. Wundt, *Gundriss der Psychologie [Outlines of Psychology]*, Leipzig, Germany: Entgelman, 1896.

[37] Bradley M. M. Lang P. J., Greenwald M. K. and Hamm A. O., "Looking at pictures: Affective, facial, visceral, and behavioral reactions," *Psychophysiology*, vol. 30, no. 3, pp. 261– 273, 1993.

[38] Gaetan Lorho, "Evaluation of spatial enhancement systems for stereo headphone reproduction by preference and attribute rating," May 2005, Audio Engineering Society.

[39] Slawomir Zielinski, Francis Rumsey, and Søren Bech, "On some biases encountered in modern audio quality listening tests-a review," *Journal of the Audio Engineering Society*, vol. 56, no. 6, pp. 427–451, June 2008.

[40] "ITU-R BT.500-10: methodology for the subjective assessment of the quality of television pictures," Tech. Rep., 2000.

[41] "ITU-R BS.1534-1: method for the subjective assessment of intermediate quality levels of coding systems," Tech. Rep., 2003.

[42] J. Faure, "Evaluation de la synthese binaurale dynamique," Tech. Rep., France Telecom, 2005.

[43] A. Gonot, *Conception et évaluation d'interfaces de navigation dans les environnements sonores 3D*, Ph.D. thesis, CNAM, Paris, 2008.

[44] P. Guillon, *Individualisation des indices spectraux pour la synthèse binaurale: recherche et exploitation des similarités inter-individuelles pour l'adaptation ou la reconstruction de HRTF*, Ph.D. thesis, Université du Maine, Le Mans, France, 2009.

[45] J. Lassalle, *Etude de l'influence de la qualité audiovisuelle sur la qualite d'expérience du spectateur: Combinaison d'indicateurs subjectifs, physiologiques et oculaires*, Ph.D. thesis, Telecom Bretagne, 2013.

[46] Cédric R. André, Jean-Jacques Embrechts, Jacques G. Verly, Marc Rebillat, and Brian F. G. Katz, "Sound for 3D cinema and the sense of presence," in *Proceedings of the 18th International Conference on Auditory Display*, Atlanta, USA, June 2012.

[47] Jenison R. Mrsic-Flogel T., King A. and Schnupp J., "Listening through different ears alters spatial response fields in ferret primary auditory cortex," *J. Neurophysiol.*, vol. 86, pp. 1043–1046, 2001.

[48] Gunther Theile, "On the naturalness of two-channel stereo sound," *The Journal of the Audio Engineering Society*, vol. 39, no. 10, pp. 761–767, Oct. 1991.

[49] L. L. Beranek, *Music, Acoustics & Architecture*, John Wiley, New York, 1962.

COMPARISON OF A 2D- AND 3D-BASED GRAPHICAL USER INTERFACE FOR LOCALIZATION LISTENING TESTS

Michael Schoeffler, Susanne Westphal, Alexander Adami, Harald Bayerlein, Jürgen Herre

International Audio Laboratories Erlangen
A Joint Institution of Fraunhofer IIS and University of Erlangen-Nürnberg
Erlangen, Germany
michael.schoeffler@audiolabs-erlangen.de

ABSTRACT

Recently, there is a trend in developing new multi-channel formats towards adding additional loudspeakers in elevated positions. While the common 5.1 surround sound system only has loudspeakers in the horizontal plane, more complex systems, such as 10.2 or 22.2, include two or more elevated loudspeakers.
When listening to music using a multi-channel playback system, the audio material has often not been produced for the used system, e.g. listening to 10.2 material while using a 5.1 surround system. In such cases, the audio material has to be down- or upmixed. Compared with listening to the original audio material, down- or upmixing affects the listening experience. The localization of sound sources is one attribute that might be affected by down- or upmixing the audio material.
In the past, some localization listening tests were conducted by using an user interface depicting a two-dimensional representation of the scene. When it comes to elevated loudspeakers, a third dimension also has to be depicted by the user interface. In this work, an experiment was conducted where participants had to locate sound sources by using two different graphical user interfaces (GUIs). The first GUI consisted of two static images of the scene: a top-view and a front-view. The other GUI had a fully adjustable 3D visualization of the scene. The main purpose of the experiment is to investigate the differences between both GUIs. This includes the time participants spend on each GUI and the difference in the responses. This work is a contribution to the development of new evaluation methods for new and existing multi-channel audio formats and renderers.

1. INTRODUCTION

A number of localization experiments were conducted to find out more about the human ability to localize sound sources. In experiments, reporting the perceived location of sound sources by pointing (with or without the extension of a body part) has been found to be the most accurate method [1]. Due to high accuracy, pointing methods were widely used in recent experiments (e.g. [2][3][4]). However, one drawback of pointing methods is that they can only be applied when localizing sound sources in the listener's field of view. As a consequence, pointing methods can not be used when the listener is not allowed to move his head. Such a condition has to be kept when localizing sound sources in the back. One application example of localizing sound sources in the back is the evaluation of down-mixers. A down-mixer is needed when one multi-channel format has to be converted into another multi-channel format with fewer channels. Especially considering down-mixes where the input multi-channel format contains elevated loudspeakers and the output format does not, a method is needed which supports reporting the localization of sound sources in all three dimensions. Such a method becomes even more important if the distance of a sound source has to be evaluated, too. Since pointing methods do not match the afore-mentioned requirements, two graphical user interfaces which enable the listener to indicate sources at any position are compared in this paper.

Our main research questions are: how accurate are the two types of GUIs, how much time is needed for reporting the localized stimuli and which variables influence the accuracy?

2. RELATED WORK

Graphical user interfaces were used in many localization tests before. Wenzel investigated the effect of increasing system latency on localization of virtual sounds by using a graphical response method [5]. The listener's head was displayed from a top view on the left-hand side of the GUI, while a front view was displayed on the right-hand side. Almost the same GUI was used in an experiment conducted by Begault et al [6]. The GUI of these experiments depicted only the listener's head and did not support reporting the distance.

Pernaux et al. tested three different reporting methods [7]. The first one used a 2D visual feedback of the listener's head. The second GUI displayed the listener's head in a three-dimensional view. The third one was similar to the second one apart from using a 3D finger pointing input instead of a computer mouse. They observed significant differences between these three reporting methods. The 2D-based and 3D-based GUIs of their experiments did not support reporting the distance of a sound source.

Martin et al. utilized a GUI displaying a top view of the scene to investigate the localization using a five-channel surround sound reproduction system [8]. The elevation and distance of a sound source was not investigated in their experiment.

Choisel and Zimmer developed a new pointing method for localizing frontal sources and compared it with a graphical response method [9]. In their experiment, only azimuth angles of sound sources in front of the participants were tested.

Yoo et al. evaluated the localization of sound sources on the horizontal plane for a wave field synthesis system [10]. Listeners reported the sound source positions using an answer sheet which contained a scheme of the scene. Two different listening positions were examined including front and back sound sources. Listeners were found out to have an average localization error of 6.1° azimuth and 9.18° elevation, while the average distance error was 0.5 m and 0.6 m in the horizontal and vertical plane, respectively.

In contrast to many localization tests conducted before, our GUI allows to report the location of a sound source in all three dimensions. By comparing two types of GUIs, the effect size of the GUIs can be measured. Except for the experiment conducted by Pernaux et al., we found no studies which investigated the difference between GUIs for the same experiment setup. Our experiment covers reporting of front, side and back sources with listeners being allowed to move their heads only slightly.

3. METHOD

3.1. Stimuli

Three different signals were used for generating the stimuli. The first signal was pink noise which was faded in and out over 500 ms. The pink noise signal was only played back during the training phase. The second signal was a sine wave with a frequency of 220 Hz and also faded in and out over 500 ms. The steady sine signal represented a sound source which is hard to localize according to Hartmann [11]. The third signal was a castanet recording. In contrast to the sine signal, the castanets recording represented a narrow sound source which is easier to localize due to its transient structure. All signals had a length of 7.8 s and were adjusted to have equal loudness by two expert listeners. The loudness was adjusted using the final experiment setup.

Six loudspeakers were used to reproduce the sine signal and castanets recording. Furthermore, one additional loudspeaker was exclusively used for the training. The loudspeakers, with positions according to Table 1, were placed in 1.5 m distance relative to the participant. We categorized the loudspeaker positions dependent on their azimuth angles to *front*, *side* and *back*. The positions were selected based on previous research about localization and taking into account sound sources in front can be localized more accurately than sources behind the listener[12]. The positions of an established multi-channel system were not used since participants familiar with surround sound might have been biased.

No.	Azimuth	Height	Category
Training	30 °	-1 cm	-
1	10 °	-1 cm	front
2	−55 °	-1 cm	side
3	120 °	-1 cm	back
4	−10 °	33 cm	front
5	55 °	33 cm	side
6	−120 °	33 cm	back

Table 1: Loudspeaker positions are relative to the listener's head (height = 120 cm). The height of the loudspeakers was measured from the loudspeakers' center.

Summarized, two different signals were played back from six different loudspeaker positions which results in a total number of twelve stimuli. For the training, a dedicated signal and loudspeaker position was used.

3.2. Participants

Thirty participants including twenty audio professionals took part in the experiment. Most of the participants were employees or students of the International Audio Laboratories Erlangen. Details about the participants are given in Table 2.

Participants		30
Audio professionals		20
Familiar with surround sound		5
Familiar with listening tests		25
Age groups [years]:	[0 − 19]	1
	[20 − 29]	19
	[30 − 39]	5
	[40 − 59]	5

Table 2: Detailed information about the participants.

3.3. Materials and Apparatus

3.3.1. Setup

The experiment took place in a soundproof listening room with room measurements (H x W x D) 256 x 452 x 455 cm. In the middle of the room, a chair and a table were placed for the participants. A 24" widescreen LCD monitor mounted on a small table was placed in front of the chair and table.

The loudspeakers were of type Focal CMS40 with measurements (H x W x D) 23.8 x 15.6 x 15.5 cm. A black-colored 360 ° masking curtain made of deco-molton was installed to veil the loudspeakers. The curtain was fixed to an aluminum ring with a diameter of 2 m which was attached to three truss stands at a height of 212 cm. The lighting in the room was adjusted such that participants could not spot the loudspeakers beyond the curtain. The masking curtain attenuated frequencies above 300 Hz by constantly 2-3 dB.

A face-tracking system was installed to prevent participants from moving their head while locating the stimuli. When participants nodded or turned their head more than 25 °, a warning message popped up and the stimulus stopped playing.

A picture of the experiment setup is shown in Figure 1.

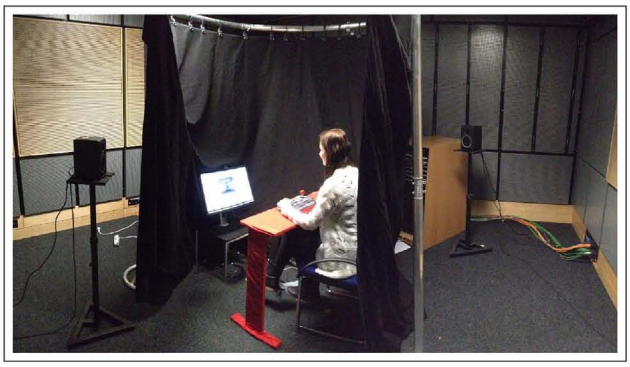


Figure 1: A picture from the setup. The masking curtain was closed during the experiment.

3.3.2. 2D-based GUI

The 2D-based GUI had two orthographic views of the same virtual scene which was a representation of the room the participants were sitting in. On the left-hand side, a top view of the scene was shown whereas a front view was presented on the right-hand side. The virtual scene contained the participant's head, a monitor, the masking curtain and a red sphere. Participants could adjust the position and size of the red sphere and thus indicate where they

localized the stimulus. The scene including all modeled objects was true to scale. A screenshot of the 2D-based GUI is depicted in Figure 2.

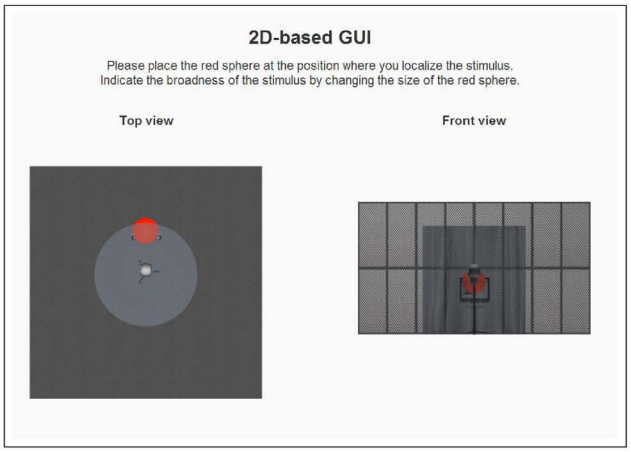


Figure 2: Screenshot of the 2D-based GUI.

3.3.3. 3D-based GUI

The 3D-based GUI was almost similar to the 2D-based GUI except for solely a single perspective view was displayed instead of two orthographic views. The virtual camera of this view was controllable by the participants to select the preferred camera views. Figure 3 shows a screenshot of the 3D-based GUI.

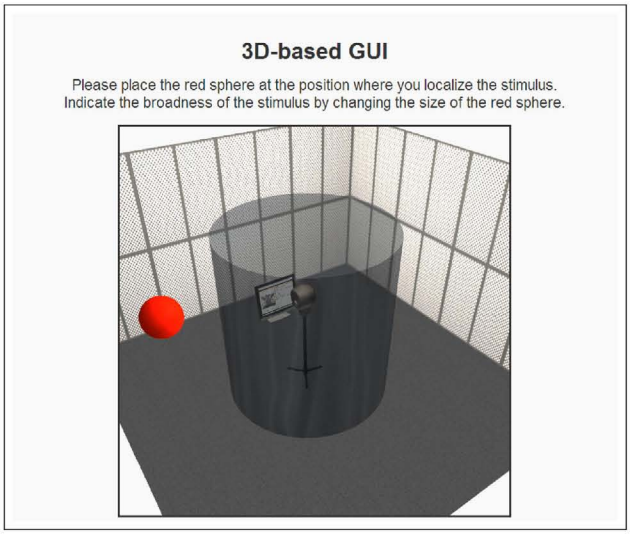


Figure 3: Screenshot of the 3D-based GUI.

3.3.4. Input Controller

For reporting the position of the red sphere, a custom-made input controller was developed. This controller offered three different types of inputs:

The camera could be moved along a sphere using an analog joystick and was always directed towards the participant's virtual representation. Zooming in and out could be done by pressing the according button next to the camera controlling joystick. Camera controls were only active while using the 3D-based GUI.

The red sphere could be moved on the horizontal plane by using a digital joystick. Each step corresponded to an accuracy of 5 cm. For moving the red sphere up- or downwards, two additional buttons were located next to the digital joystick. By pressing a button once, the red sphere moved 5 cm. The size of the sphere could be adjusted by another two buttons. The sphere controls were active while using the 2D-based as well as the 3D-based GUI.

Furthermore, two buttons for playing back the stimulus (play button) and completing the response (next button) were located below the camera and sphere controls. When a stimulus was already playing, pressing the play button had no effect.

Developing an own custom input controller allowed us to design an individual arrangement of buttons which is easy to understand. E.g. if a keyboard had been used, participants might have spent more time on learning the relevant keys and their function. In Figure 4, a picture of the input controller is shown.

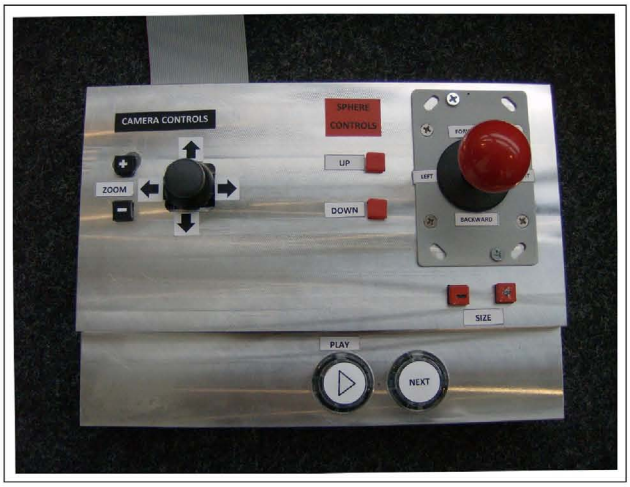


Figure 4: The input controller.

3.4. Procedure

The experiment had a subject-within design. All participants had to localize all twelve stimuli using both GUIs. In total, each participant gave twenty-four responses.

All participants were blindfolded and guided by an experimenter to the chair in the middle of the room. Blindfolding the participants assured that they can not spot loudspeakers while entering the room. After removing the blindfold, participants were instructed to always keep their heads straight towards the monitor during the experiment. Furthermore, they were informed that their faces would be tracked by a face-tracking system to verify that they would be looking towards the monitor. The experimenter then left the room and all subsequent instructions were displayed by the experiment software.

At the beginning of the test, participants had to fill out a questionnaire. They were asked whether they regularly listen to surround sound, whether they are an audio professional, if they are familiar with listening tests and to which age group they belong. The questionnaire was followed by some general instructions: The

participants were again reminded that they had to localize stimuli and are not allowed to turn their head. The general instructions announced that a 2D- and a 3D-based GUI would be used for reporting the location of the stimuli.

It was randomly chosen which GUI was initially presented to the participants. Before the participants could report the stimuli locations, they had to read the detailed instructions. These contained a brief description of the GUI, how the input controller worked and that they are asked to localize the stimuli. For the 3D-based GUI, additional information about moving the camera was included in the detailed instructions. After reading the instructions, the participants had to undertake training in which they were asked to place the red sphere at the position where they localize the stimulus. Additionally, they were asked to indicate the broadness of the stimulus by changing the size of the red sphere. The tutorial could only be finished if every control element was used at least once (play button, position and size of the red sphere). In the tutorial for the 3D-based GUI, participants also had to move the camera. Afterwards, participants had to localize twelve stimuli using the present GUI. When they finished reporting all stimuli locations, the same procedure was applied for the second GUI.

At the end of the experiment, the participants had to fill out a questionnaire about how they got along with each GUI.

4. RESULTS

Completing the experiment took 19 minutes (SD[1] $= 6$) on average. To analyze the accuracy of both GUIs, we define the localization error as the euclidean distance between the reported position $\mathbf{r}$ and the actual loudspeaker position $\mathbf{l}$:

$$LocError = ||\mathbf{r} - \mathbf{l}||_2, \tag{1}$$

where bold-faced letters represent vectors with $\mathbf{v} = [v_x, v_y, v_z]^T$. The mean of $LocError$ was 82 cm ($SD = 68$ cm) for all stimuli. The 2D-based GUI had a $LocError$ mean of 83 cm ($SD = 68$ cm) for all stimuli. The 3D-based GUI had a mean $LocError$ of 82 cm ($SD = 68$ cm). The effect of the GUI on $LocError$ was not significant at the $p < .05$ level [$F(1, 717) = 0.094$, $p = .759$]. Levene's test indicated equal variances for $LocError$ ($F = 0.20, p = .655$). In Table 3, detailed information about $LocError$ is given.

	2D		3D		both	
	M	SD	M	SD	M	SD
front	55	52	60	53	58	52
side	66	47	71	65	68	57
back	129	76	114	71	121	74
all	83	68	82	68	82	68

Table 3: Average Localization errors in cm for both stimuli. The table is segmented by the GUI type and the loudspeaker category.

The distance of a stimulus is dependent, among other factors, on its loudness [13]. As the loudness was only subjectively adjusted by two expert listeners, a normalized localization error is calculated. The normalized localization error excludes the depth distance of the distance between reported position and loudspeaker position. The normalized reported position is defined as:

$$\mathbf{r}_{\text{Norm}} = \frac{\mathbf{r}}{|\mathbf{r}|}|\mathbf{l}|. \tag{2}$$

The normalized reported position has the same direct distance to the listener as the loudspeaker and is used for calculating the normalized localization error:

$$LocError_{\text{Norm}} = ||\mathbf{r}_{\text{Norm}} - \mathbf{l}||_2. \tag{3}$$

The mean of the normalized localization error was 70 cm ($SD = 62$). The 3D-based GUI ($M = 67$ cm, $SD = 59$) turned out to be more accurate than the 2D-based GUI ($M = 73$ cm, $SD = 65$). According to a repeated measures ANOVA, the difference between the two GUIs was not statistically significant at the $p < .05$ level [$F(1, 717) = 2.295, p = .13$]. Levene's test indicated equal variances for $LocError\text{Norm}$ ($F = 1.710, p = .192$). In Table 4, all values for the normalized localization error can be found.

	2D		3D		both	
	M	SD	M	SD	M	SD
front	46	55	40	49	43	52
side	57	42	57	42	57	42
back	119	69	104	66	111	68
all	74	65	67	59	70	62

Table 4: Normalized localization errors in cm for both stimuli. The table is segmented by the GUI type and the loudspeaker category.

A linear regression model with mixed effects was calculated to analyze the influences on the normalized localization error in more detail. The participants were defined as random factor (*participants_id*). The fixed factors were the type of GUI (*GUI*), the loudspeaker category (*category*), the signal (*signal*), the total time spent on the GUI type (*time_GUI*), the time spent on the training for the GUI type (*time_training*) and if the response was given when the GUI was chosen last. (*GUI_last*). The results of the fitted model are described in Table 5.

Coefficient	Value	Std. Error	t-value	p-value
Fixed Effects:				
(Intercept)	34.25	7.9	4.34	.000
GUI = 3D	-3.43	4.58	-0.75	.455
category = side	13.49	4.6	2.92	.004
category = back	68.05	4.6	14.71	.000
signal = sine	36.32	3.78	9.61	.000
time_GUI	-1.61	1.00	-1.62	.105
time_training	4.34	2.57	1.69	.091
GUI_last = true	-5.87	4.29	-1.37	.172
Random Effects:				
participants_id				
	(Intercept)	Residual		
StdDev:	8.52	50.69		

Table 5: Linear regression model with mixed effects for the normalized localization error. The table is segmented by the GUI type and the loudspeaker category.

The normalized localization error can also be expressed as the

[1] M = mean, SD = standard deviation.

elevation and azimuth error of the normalized reported position ($EleError_{\mathrm{Norm}}$ and $AziError_{\mathrm{Norm}}$). The mean of $EleError_{\mathrm{Norm}}$ was $10\,^{\circ}$ ($SD = 34$). As indicated by the linear regression model of the normalized localization error, the signal type had an influence on the elevation and azimuth errors. The average normalized elevation error of the castanets recording was $9\,^{\circ}$ ($SD = 7$). When sine wave was played back the average normalized elevation error increased ($M = 11\,^{\circ}, SD = 8$). If the normalized elevation errors are analyzed for each loudspeaker category, the castanets recording resulted in lower errors for each category (front: $M = 8\,^{\circ}, SD = 6$; side: $M = 9\,^{\circ}, SD = 6$; back: $M = 10\,^{\circ}, SD = 8$). The sine wave resulted in much higher average normalized elevation errors (front: $M = 11\,^{\circ}, SD = 9$; side: $M = 11\,^{\circ}, SD = 7$; back: $M = 10\,^{\circ}, SD = 7$). All normalized elevation errors for both stimuli are described in Table 6.

	2D		**3D**		**both**	
	M	SD	M	SD	M	SD
front	10	8	9	7	10	8
side	10	7	9	7	10	7
back	11	7	9	8	10	8
all	11	8	9	7	10	7

Table 6: Elevation errors of the normalized reported position in degrees. The table is segmented by the GUI type and the loudspeaker category.

The mean of $AziError_{\mathrm{Norm}}$ was $24\,^{\circ}$ ($SD = 30$). As expected, the average normalized azimuth errors widely differed for each signal type. The average normalized azimuth error of the castanets recording was $16\,^{\circ}$ ($SD = 21$). The average normalized azimuth error increased when the sine wave was played back ($M = 32\,^{\circ}, SD = 34$). The castanets recording resulted in smaller errors for all categories (front: $M = 5\,^{\circ}, SD = 4$; side: $M = 13\,^{\circ}, SD = 13$; back: $M = 30\,^{\circ}, SD = 29$). As expected, the sine wave resulted in larger average normalized azimuth errors (front: $M = 21\,^{\circ}, SD = 38$; side: $M = 22\,^{\circ}, SD = 19$; back: $M = 54\,^{\circ}, SD = 31$). All normalized azimuth errors are described in Table 7 for both stimuli.

	2D		**3D**		**both**	
	M	SD	M	SD	M	SD
front	14	29	12	27	13	28
side	18	17	18	17	18	17
back	46	33	39	31	42	32
all	26	31	23	28	24	30

Table 7: Azimuth errors of the normalized reported position in degrees. The table is segmented by the GUI type and the loudspeaker category.

The differences of localization errors between the sine wave and the castanets recording are confirmed by the reported broadness of the two stimuli. The average radius of the red sphere was 23 cm($SD = 7$) when the castanets recording was played back. For the sine wave stimulus, the average reported radius of the sphere was $M = 45\,cm$ ($SD = 29$).

Reporting twelve stimuli for the 2D-based GUI took the participants in average 5 minutes ($SD = 2$). For the 3D-based GUI

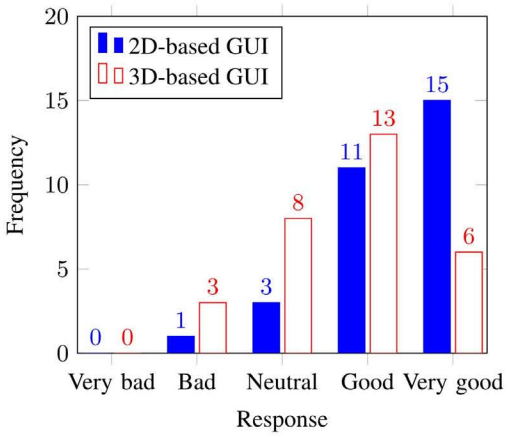


Figure 5: Frequencies of participants' responses about they got along with each GUI.

8 minutes ($SD = 3$). According to a repeated measures ANOVA, the difference of the time participants spent on each GUI was statistically significant at the $p < .05$ level [$F(1, 57) = 17.99$, $p = .000$]. Levene's test indicated equal variances for $LocError$ (F = 4.01, p = .04997).

At the end of the experiment, the participants were asked how they got along with each GUI. The possible answers were: "Very bad" (=1), Bad (=2), "Neutral" (=3), "Good" (=4) and "Very good" (=5). In Figure 5, the frequency of the answers are shown. The mode of the 2D-based GUI was Very Good and for the 3D-based it was Good. A cumulative link model (Table 8 supported the information provided by the listeners that they got along better with the 2D-based GUI. The 3D-based GUI type (*GUI*) had a significant negative effect on the participants' answers ($F = -1.34, p = .008$). If a participant used a type of GUI last (*last*), it had a non-significant positive effect on the answer ($F = 0.46, p = .344$).

Coefficient	Estimate	Std. Error	z-value	p-value
GUI = 3D	-1.34	0.51	-2.64	.008
last = true	0.46	0.49	0.95	0.344
Threshold coefficients:				
	Estimate	Std. Error	z-value	
$Bad\|Neutral$	-3.28	0.67	-4.91	
$Neutral\|Good$	-1.67	0.50	-3.37	
$Good\|Very\ good$	0.25	0.44	0.57	
Number of observations: 60				
Cragg and Uhler's pseudo R^2: 0.14				

Table 8: Logit cumulative link model of the response about how participants got along with each GUI.

5. DISCUSSION

Participants were much faster in reporting the location when using the 2D-based GUI. This was expected since the 2D-based GUI showed two views at the same time. To have a similar perspective

with the 3D-based GUI, the camera had to be moved which took some time. Some participants reported the two views of the 2D-based GUI were perfectly good for them. When using the 3D-based GUI, they consecutively moved the camera such that a front and top view were shown. Another reason for the time differences might be due to the 3D-based GUI was something new and fun to use as some participants reported. These participants spent a bit more time on the 3D-based GUI just for playing around with the camera. Therefore, for listening tests with many items, a 2D-based GUI with multiple views should be used instead of a 3D-based GUI with a virtual camera.

Participants reported getting along better with the 2D-based GUI. This was expected since the 3D-based GUI needs the camera to be adjusted which increases the complexity. The 2D-based GUI already showed two views which enabled the user to monitor all three dimensions. The red sphere was controlled by a joystick relative to the virtual participant, even if the camera was moved. Some participants reported that they would have expected the red sphere to move relative to the camera (e.g. pressing the joystick forward moves the red sphere away from the camera).

Localization errors of the participants' responses were slightly smaller when the 3D-based GUI was used. However, the differences in the localization error and normalized localization error were not significant. Nevertheless, the similar results are interesting, considering that participants reported that they got along much better with the 2D-based GUI. By the linear regression model, it could be revealed that for predicting the localization error other effects are much more relevant than the GUI. The effect size of the GUI type was small in the model and also not significant. Loudspeaker position and the signal type influenced the localization most. This was expected since these effects are known from established research. There are non-significant indications that training is important for reporting the localization by graphical user interfaces. When the GUI was last used, the localization error was reduced according to the model.

The reported azimuth angle of the normalized location error turned out to be accurate when the castanets recording was played back by the front loudspeakers. The average normalized azimuth error was $5\,^\circ$ which is close to results achieved by other localization methods. E.g. Haber et al. reported that the average localization errors of nine different methods ranged from $+3.5\,^\circ$ to $-5.2\,^\circ$ for front loudspeaker positions[1]. In the experiment of Yoo et al. the average azimuth error was $-5\,^\circ$ and $2.3\,^\circ$ for two tested front loudspeaker positions ($-30\,^\circ$ and $+30\,^\circ$)[10].

6. CONCLUSION

A method for reporting the location of sound sources in all three dimensions was presented. The method was evaluated by conducting an experiment with two different types of GUIs: A 2D-based GUI and a 3D-based GUI. The 2D-based GUI had an average localization error of 83 cm and turned out to be the less time-consuming and more convenient choice. The 3D-based GUI was slightly more accurate and had an average localization error of 82 cm. The analysis of the experiment results revealed that the used GUI had only a small effect on the localization error. The signal type and loudspeaker position played a much more important role. For front loudspeaker positions, both GUIs resulted in an average normalized azimuth error of $5\,^\circ$ when a castanets recording was played back.

7. REFERENCES

[1] L. Haber, R. N. Haber, S. Penningroth, K. Novak, and H. Radgowski, "Comparison of nine methods of indicating the direction to objects: data from blind adults.," *Perception*, vol. 22, no. 1, pp. 35–47, Jan. 1993.

[2] M. Frank, L. Mohr, A. Sontacchi, and F. Zotter, "Flexible and Intuitive Pointing Method for 3-D Auditory Localization Experiments," in *Audio Engineering Society Conference: 38th International Conference: Sound Quality Evaluation*, 2010.

[3] H. Wierstorf, A. Raake, and S. Spors, "Localization of a Virtual Point Source within the Listening Area for Wave Field Synthesis," in *Audio Engineering Society Convention 133*, 2012.

[4] T. Ashby, R. Mason, and T. Brookes, "Head Movements in Three-Dimensional Localization," in *Audio Engineering Society Convention 134*, 2013.

[5] E. M. Wenzel, "Effect of increasing system latency on localization of virtual sounds," in *Proceedings of the Audio Engineering Society 16th International Conference on Spatial Sound Reproduction*, 1999, pp. 42–50.

[6] D. R. Begault, E. M. Wenzel, and M. R. Anderson, "Direct comparison of the impact of head tracking, reverberation, and individualized head-related transfer functions on the spatial perception of a virtual speech source.," *Journal of the Audio Engineering Society. Audio Engineering Society*, vol. 49, no. 10, pp. 904–916, Oct. 2001.

[7] J. Pernaux, M. Emerit, and R. Nicol, "Perceptual Evaluation of Binaural Sound Synthesis: the Problem of Reporting Localization Judgments," in *Audio Engineering Society Convention 114*, 2003.

[8] G. Martin, W. Woszczyk, J. Corey, and R. Quesnel, "Sound Source Localization in a Five-Channel Surround Sound Reproduction System," in *Audio Engineering Society Convention 107*, 1999.

[9] S. Choisel and K. Zimmer, "A pointing Technique with Visual Feedback for Sound Source Localization Experiments," in *Audio Engineering Society Convention 115*, 2003.

[10] J. Yoo, J. Seo, H. Shim, H. Chung, K. Sung, and K. Kang, "Subjective Listening Experiments on a Front and Rear Array-Based WFS System," *ETRI Journal*, vol. 33, no. 6, pp. 977–980, 2011.

[11] W. M. Hartmann, "Localization of sound in rooms," *The Journal of the Acoustical Society of America*, vol. 74, no. November 1983, pp. 1380–1391, 1983.

[12] S. Carlile, P. Leong, and S. Hyams, "The nature and distribution of errors in sound localization by human listeners.," *Hearing research*, vol. 114, no. 1-2, pp. 179–96, Dec. 1997.

[13] G. von Békésy, "The moon illusion and similar auditory phenomena," *The American journal of psychology*, vol. 62, no. 4, pp. 540–552, 1949.

Proc. of the EAA Joint Symposium on Auralization and Ambisonics, Berlin, Germany, 3-5 April 2014

AN ANECHOIC AUDIO CORPUS FOR ROOM ACOUSTICS AND RELATED STUDIES

Antti Kuusinen

Virtual Acoustics Research Group, Department of Media Technology,
Aalto University School of Science,
Espoo, Finland
antti.kuusinen@aalto.fi

ABSTRACT

Anechoic or semi-anechoic instrument recordings are readily available for academic purposes on a few different sites online. Anechoic recordings are commonly used in auralizations, which today practically means convolving recordings with simulated or measured room impulse responses. Besides the possibility of being used as such, these recordings offer other possibilities for the generation of test stimuli. Many studies, such as, studies on auditory distance perception or source separation, would benefit from available experimental materials which would not be strictly musical but could still be linked to the perception of musical stimuli. The goal of the current investigation is to develop a procedure for generating such materials, i.e., an anechoic audio corpus which can be used in the future investigations of room acoustics and in related fields. Moreover, the aim is to provide a framework for further development of processes where a large number of stimuli can be generated in a systematic way. In this study, the proposed framework is instantiated by producing two sets of stimuli by either directly segmenting anechoic music or randomly combining different segments of anechoic instrument tracks. Music information retrieval (MIR) approach is used to calculate 14 musical features of the generated sets of stimuli. Principal component analysis is used to analyse the sample spaces enabling the experimenter to select a small number stimuli with desired characteristics. The benefits and drawbacks of this stimuli generation approach including some important theoretical underpinnings of experimental design are also discussed.

1. INTRODUCTION

Auralizations are made by convolving audio signals with simulated or measured impulse responses. Making convolutions commonly require anechoic audio materials which are free from extraneous reflected sounds - especially in research of room acoustics. In the studies of music performance spaces, anechoic instrument recordings or synthesized instrument sounds are natural choices as source signals. Anechoic or semi-anechoic instrument recordings are readily available for academic purposes on a few different sites online, e.g., [1, 2]. Also commercial releases, such as, Vienna Symphonic Library [3] exist. Such recordings are usually of either single notes with various characteristics (steady, vibrato, pizzicato etc.) and/or recordings of musical sequences or excerpts of compositions with various styles and with separate tracks for each instrument or instrument section [4].

Besides being invaluable for research as such, these recordings offer other possibilities for the generation of stimuli. Many studies, such as, studies on auditory distance perception or source separation, would benefit from available experimental materials which would not be strictly musical but could still be linked to the perception of musical stimuli. Also studies where 'ecologically valid' musical stimuli, that is, stimuli which are unequivocally musical are required, could benefit from contrasting stimuli which would not be strictly musical. While the ecological validity of the test stimuli is one of the main reasons for employing commercial recordings and well known compositions, usually the recording and production chain is not well described, what is a major drawback to using such releases in scientific work. In addition, there are many situations where the selection of test stimuli unwarrantedly restricts the experimenter from making more generalizable inferences from the results. This is true especially in situations where musical stimuli are used in the subjective evaluation of "treatments" such as different signal processing algorithms, auralization methods, or different room acoustical conditions.

Consider a simple case where a researcher investigates the performance of a few reverberation algorithms with a listening experiment. Let's say that the number of algorithms, i.e., treatments, to evaluate is 6. The perceptual task is to indicate for each sound how "natural" the reverberation is on a 20-point scale. Then he or she chooses four anechoic source signals with different characteristics which he/she thinks is a representative sample of the population of sounds to which the algorithms would be used. Listening experiment is conducted in blocks, where the algorithms are compared in parallel with each source signal. A number of assessors participates in the experiment and data is analysed with the analysis of variance (anova).

The different sources of variance, that is, factors, must be specified in setting up an anova model. Moreover, one must specify whether each factor is treated as "fixed" or "random". In our example, the main factors are the treatments (the algorithms), the source signals and the individuals. Whether to treat these factors as fixed or random is essentially based on the nature of inferences one wishes make to about the possible results. Treating a factor as fixed indicates that the different levels of that factor are exhaustive of the population that the factor represent, that is, the inferences are restricted to these particular treatments. Clearly, the treatments are considered as fixed in our current example as the objective is to evaluate the perceptual differences in this particular set of algorithms. In contrast, treating a factor as random indicates that the factor levels are drawn randomly from the population of interest and possible inferences made from the results generalize to that population. In other words, the differences in these levels of the factor are not of particular interest, but rather the effect of changes between the factor levels in general. In our example, individuals can be considered as being randomly selected from a population and thus, as a random factor.

Treating the algorithms as a fixed factor and the individuals

as a random factor is quite straightforward, but the source signal might be considered as a fixed or a random depending on the assumptions and inferences the experimenter wishes to make. If treated as a fixed factor, the conclusions are made specifically about these source signals. In other terms, the algorithms are studied only with respect to these particular source signals making the experiment into a 'case' study. It is of course possible, that as a case study, the possible inferences may have apparent theoretical implications which account for a more general discussion about the results. In addition, treating a factor as fixed also has the benefit for statistical analysis to reveal smaller effects, which is apparent by considering the sources of variances included in the error component of a model. A detailed discussion about these aspects is outside the scope of this paper (see more in, e.g., [5] or [6]), but basically treating a blocking factor, in our case the source signal, as random, adds uncertainty (i.e., wider confidence intervals) to the analysis of treatment means. However, if the treatments are still significantly different from each other when the source signal is treated as a random factor, the inferences from these results can be generalized to the population from which these source signals are randomly chosen.

Thus, in our example, treating the source signal as a random factor means that the experimenter can generalize the possible results about the performance of the reverberation algorithms outside the set of source signals used in the experiment. This is clearly desirable in many situations, but, this also implies that the source signals should be a representative sample randomly drawn from a population of possible source signals. Clearly, this population can be stated as infinite and impossible to specify, so that the experimenter may well choose the source samples on the grounds of his/her best knowledge and intuition. Unfortunately, subjective knowledge and intuition are often rather problematic bases for scientific work. Thus, here I propose a simple way to help researchers to select source signals by first producing a large set of samples, that is a population of stimuli, which can then be randomly sampled as desired. Of course, the issue of the representativeness of this population to an infinite sound space still remains, but the major advantage is the knowledge about the population of stimuli to which the results should be generalizable. Keeping these considerations in mind, the next section provides some further contextual aspects and reasons for why such processes are needed.

2. SELECTING STIMULI FOR LISTENING EXPERIMENTS

As discussed above, the information the experimenter expects to extract from the results determines the design of the experiment - including the selection of the test stimuli. In different fields of audio research different types of stimuli are typically used. For instance, in psychoacoustic research often used stimuli are noises and pure tones with various derivatives and variations (see e.g. [7], p. 2). In speech perception studies a natural choice is to use samples of speech. In music related studies some excerpts of music is a common choice. The performances of reproduction systems and perceptual coding algorithms are commonly studied with stimuli which is known to be particularly revealing about a certain effect; e.g., sound of castanets is typically used to reveal about unwanted pre-echos in perceptual coding. A simplified schematic of different stimulus types is presented in Fig. 1. Of course any combination of different types of stimuli is possible if needed and in fact, often there is no clear distinction between stimulus types. In Fig. 1 the miscellaneous stimuli are presented by a big surrounding circle while the more specific types of stimuli are represented by the circles.

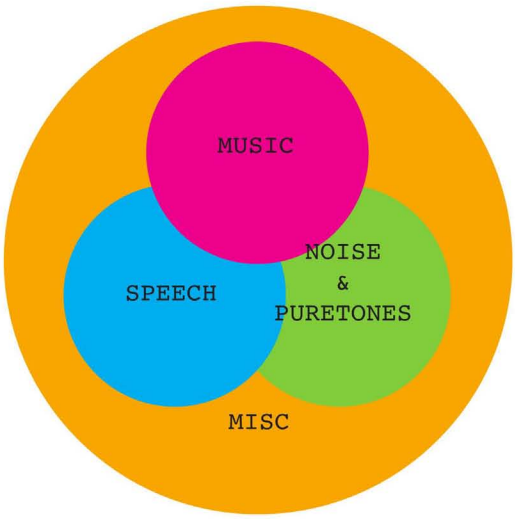


Figure 1: *A schematic of different types of stimuli.*

There are many fields of audio research, such as room acoustics, where practically any stimulus type can be used. The generation and manipulation of noises and pure tones have become every day practice with modern digital signal processing techniques and now these types of stimuli can be quite easily obtained when needed. Considering speech related studies, the situation is not as straightforward, but there are readily available speech samples which have been widely employed in speech intelligibility and related studies. The Coordinate Response Measure (CRM) corpus [8] was originally developed for a particular speech recognition task but has been extended [9] to enable studies also in audio-visual domain. The many advantages of a stimulus set, which is widely adopted and used in a research field, include: facilitated experimental design, easier comparison and cross-validation of studies performed in different laboratories and possibly more efficient "steering" of research with the emergence of relevant research problems.

Studies of music and related aspects are performed in a wide variety of research fields ranging from neuroscience, psychology, aesthetics and musicology to recording techniques, signal processing and room acoustics. Test stimuli are commonly excerpts of music which are most often selected by the researcher. Regarding research in music and emotion, for instance, Eerola et al. [10] reviewed 170 studies and report that the stimuli selection method has been almost entirely researcher-driven (96 %) in a sense that the choices were based on either researcher's intuition (33 %), on a pilot study (8 %), on a selection by a group of experts (6 %) or on some previous study (9 %). Although these numbers might be relevant only in that particular field, there is little doubt that researchers' knowledge and intuition would be the main determinants in the selection of test material in most experiments where musical stimuli are required, independent of the research field. Moreover, in the field of neuroscience the lack of a systematic approach to selecting musical stimuli has been argued to be one possible reason for inconsistencies between studies [11].

Based on this discussion and a few statistical principles mentioned before, it is clear that systematic approaches to stimulus selection would be beneficial for various fields of audio research. Here, one such method is proposed to help researchers in the sample selection process, as well as to stipulate critical discussion on the topic.

3. ANECHOIC AUDIO CORPUS

In order to enable a random sampling of source stimuli and to provide a framework for sample selection and future developments of similar processes, an audio corpus and a simple stimuli production method is proposed. The stimuli production method is based on the segmentation of pre-recorded anechoic material and making combinations of these segments. In particular, the aim of the current investigation is to generate a population of test materials which would allow random sampling, and where the stimuli 1) would be consisted of anechoic sounds of real instruments. 2) would include both musical stimuli as well as stimuli which are not distinctively musical as in the terms of a melody or a harmony, 3) would enable control over the acoustic ("musical") features [12] (dynamics, rhythm, timbre, pitch and tonality) of the selected stimuli and 4) could be used in a wide range of listening experiments. Moreover, the aim is to develop a systematic stimuli selection procedure, which can be used in conjunction with the experimenter's knowledge and intuition. It is worth to mention that the current work does not attempt to produce a selection procedure for "ecologically valid" musical signals although the first method proposed below also fullfils this criterium to a certain extent. The aim is to provide a framework for further development of processes where (random) stimuli can be generated and selected in a systematic way.

3.1. Background

The starting point for the current work is the anechoic symphony orchestra recordings made by Pätynen *et al.* [4]. These recordings consist of the following excerpts:

- W. A. Mozart: Aria of Donna Elvira from the opera Don Giovanni (3 min 47 s)
- L. v. Beethoven: Symphony no. 7, I movement, bars 1-53 (3 min 11 s)
- A. Bruckner: Symphony no. 8, II movement, bars 1-61 (1 min 27 s)
- G. Mahler: Symphony no. 1, IV movement, bara 1-72 (2 min 12 s)

The Fig. 2 represents the chromagrams of the music pieces. Each instrument has been recorded (48 kHz, 16 bits) separately in an anechoic chamber. Also some editing and noise reduction have been applied on the separate instrument tracks (see details of the recording process in [4] and [13]). Clearly, one could also use any recordings, such as, single notes, but in order to preserve musical characteristics, which occur naturally in played music, such as transitions, these recordings were thought to be the most appropriate for the current investigation.

The characterization of the stimulus space is inspired by recent developments in the field of music information retrieval (MIR), where acoustic features calculated directly from audio signals are used for various purposes, particularly for automatic classification

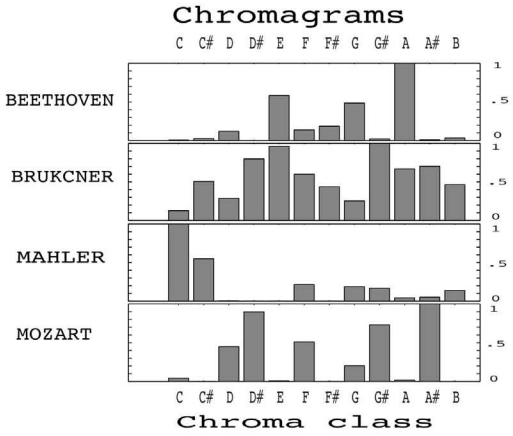


Figure 2: *Chromagrams of the original excerpts of symphonic music.*

tasks (genre, mood etc.) (see e.g. [14] for review). The extracted features are thought to represent the essential musical characteristics in the signal and to correspond to human perception to some extent. Commonly, the features represent the musical dimensions of pitch, dynamics, rhythm, timbre and tonality. Previously feature extraction and classifier training were performed in Marsyas [15] framework with WEKA machine learning software but more recently also a Matlab toolbox (MIRtoolbox [12]) has been developed. Matlab and MIRtoolbox are used in the current investigation for acoustic feature extraction. Otherwise, the data analysis is performed in R statistical programming language.

3.2. Acoustic features

A set of 14 acoustic features were selected to characterize the musical properties of the stimuli. Regarding the computational time needed for calculating each feature for each individual sample, the following features were considered covering the main musical aspects including timbre, dynamics, tonality and rhythm:

- Timbral:
  - zero-crossing rate (sign change in the signal per second),
  - spectral roll-off (a cut-off frequency below which lies 85 % of total energy),
  - brightness (energy above 1500 Hz),
  - spectral flatness,
  - spectral entropy,
  - roughness.
- Dynamics:
  - root-mean-square energy (frame length of 50 ms, 50 % overlap)
  - low to high energy ratio (% of frames with less than average energy in the segment)
- Tonality:
  - spectral flux (distance between the spectrum of successive frames of 50 ms, 50 % overlap),
  - key clarity,
  - mode (minor (-1) – major (+1)),
- Rhythm:
  - Tempo

The values were averaged over the duration of each segment to obtain single values for subsequent analysis. The calculation of these features was performed by using the default configurations defined in MIRtoolbox [12] because there was no reasons were found to change these settings. However, it is acknowledged that the settings, such as, the durations of temporal windows and the percentages of overlap affect the calculations, and closer investigation to these aspects would be beneficial for future work.

3.3. Stimulus production

To make the stimulus production process more tangible, the length of the stimuli to produce was arbitrarily set to five seconds. Of course, in real applications the length of the stimuli would be determined by the experimental design and the context and objectives of a study. Two different stimulus production methods are proposed. The first method produces "ecologically valid" musical stimuli, but is restricted by the available musical variations in the original music pieces as well as the desired length of stimuli. The second method produces a larger set of random stimuli, which might be musically questionable but enables control over various other aspects of the stimulus set, such as the number of instruments. Both of these methods enable a random sampling of the produced stimulus population. In addition, the characterization of these stimulus populations by musical features combined with principal component analysis allows one to include covariates in statistical models, thus enabling the evaluation of these effects in the results. The principal component analysis used below is performed on the correlation matrix.

3.3.1. Method 1.

First method is inspired by a method represented by Alluri *et al.* [16]. The aim is to select five second segments from the original music pieces, that would capture the range of musical variations embedded in these music pieces. The original instrument recordings are combined into one channel and 5 second segments with 1 second hop size (80 % overlap) are extracted. Including all music pieces, the total number of segments obtained and subsequently analyzed was 621. Musical features are calculated for each segment and the features containing values for shorter temporal windows are averaged over the 5 second period. Then, principal component analysis (PCA) is performed in order to reduce the dimensionality of the feature space, and to reveal associations between different features. Principal components with eigenvalues larger than 1 are selected, and varimax rotation is performed with the selected components in order to clarify the structure and interpretation of the reduced space. Finally, this reduced space can be used to select sets of samples in various ways.

Here, the averaged values of the features of the five second segments from all recordings were combined to the same data matrix which was subjected to PCA. The first 5 principal components explained 81 % of the total variance and were retained. Table 1 presents the feature loadings these PCs after varimax rotation. Althought the interpretation of principal components should be validated by a perceptual experiment, as in Alluri *et al.* [16], the PC loadings indicate that the first PC is associated with the timbral properties, the second to the sensory dissonance or consonance, the third to the dynamics and PC4 and PC5 to the tonal characteristics of the samples.

Table 1: *Feature loadings on five principal components after varimax rotation. Features were calculated for 621 stimuli produced by extracting 5 second segments of anechoich music (see text for details).*

	PC1	PC2	PC3	PC4	PC5
% of var.	32 %	14 %	13 %	13 %	13 %
Zerocross	**0.74**	-0.09	0.37	-0.04	-0.12
Rolloff	**0.77**	0.42	0.14	0.31	0.10
Brightness	**0.89**	-0.05	-0.02	-0.07	-0.19
Flatness	**0.80**	0.51	0.05	0.20	0.14
spcEnt	**0.91**	-0.17	-0.06	-0.02	0.13
spcCent	**0.85**	0.40	0.17	0.17	0.05
Roughness	0.07	**-0.87**	0.00	0.21	0.16
Spread	0.52	**0.71**	0.04	0.34	0.20
Rms	-0.31	0.10	**-0.85**	-0.02	0.16
Lowenergy	-0.02	0.14	**0.83**	-0.07	0.05
spcFlux	-0.11	0.05	-0.39	**-0.80**	0.06
Keycla	0.01	-0.01	-0.18	**0.79**	0.16
Mode	0.02	-0.04	-0.06	0.00	**0.92**
Tempo	-0.05	-0.03	0.32	-0.47	0.24

3.3.2. Method 2

In contrast to the method described above, here the aim was to produce a large set of stimuli, which consist of anechoic instrument sounds but are not predetermined or deliberately composed in musical terms. First, each anechoic instrument recording is segmented at silent periods determined by root-mean-square energy in 500 ms frames and 100 ms hop size. The segments from each music piece were grouped into instrument banks. In the present case 15 different instrument banks were obtained, reflecting the instrumentation of the original compositions. The respective instruments are bassoon, cello, clarinet, contrabass, French horn, flute, oboe, timpani, trombone, trumpet, tuba, viola, 1. violins and 2. violins.

To produce a large randomized set of 5 second samples, one sample of each instrument bank was randomly selected and combined with randomly selected samples of other instrument banks. Here, only segments shorter than 5 seconds were used. The temporal positions of the selected segments were randomly varied inside the five second sample to avoid 'stacking' the sounds to the beginning of the samples.

In this randomization procedure, it is also possible to control the number of instruments included in the random combinations. In the current implementation, the composition of instruments in a classical orchestra was used with an added trombone and a percussion instrument. In this orchestration the number of instruments, i.e., the number of randomly selected samples of the instrument banks was as follows: 2 flutes, 2 oboes, 2 clarinets, 2 bassoons, 2 French horns, 2 trumpets, 2 timpani, 10 1. violins, 8 2. violins, 6 violas, 4 cellos and 2 double bass. 1000 random combinations of the segmented instrument samples in this orchestration were produced. Features were calculated for each combination separately and average values were extracted. Like in the previous method, this feature set was analysed using PCA and varimax. Again, the first 5 principal components were retained and together they explained 82 % of the total variance. However, in contrast to the first method where the explained variances were more equally dis-

tributed between the PCs, we now observe that the PC1 explains as much as 43 % of the total variance even when the varimax rotation tends to make the explained variances more equal. The feature loadings on these PCs are tabulated in Table 2. The loadings indicate that again the first component is associated to timbral features, but now also dissonance related aspects are included. The second component is associated with dynamics, the third with tonality, and interestingly mode and tempo parameters uniquely characterize the fourth and fifth component, respectively.

Table 2: *Feature loadings on five principal components after varimax rotation. Features were calculated for 1000 stimuli produced by randomly combining the 5 second segments of anechoic instrument sounds (see text for details).*

	PC1	PC2	PC3	PC4	PC5
% of var.	43 %	15 %	9 %	7 %	7 %
Zerocross	**0.79**	0.22	-0.13	0.03	-0.04
Rolloff	**0.93**	-0.11	-0.04	-0.04	0.00
Brightness	**0.87**	0.16	-0.17	0.01	-0.05
Flatness	**0.93**	-0.28	0.02	-0.05	-0.03
spcEnt	**0.93**	-0.03	0.13	0.01	-0.03
spcCent	**0.98**	-0.06	-0.08	-0.04	-0.03
Spread	**0.69**	**-0.53**	-0.08	-0.07	-0.01
Roughness	**0.70**	0.29	0.14	0.11	0.05
Rms	-0.02	**0.90**	0.06	0.00	0.00
Lowenergy	-0.05	**-0.82**	0.12	0.03	-0.01
spcFlux	0.11	0.10	**0.91**	-0.03	-0.02
Keycla	0.31	0.29	**-0.58**	-0.16	-0.14
Mode	0.01	-0.02	0.05	**0.99**	0.01
Tempo	-0.03	0.01	0.05	0.01	**0.99**

3.4. Selection of samples

The sample space characterized by musical features and ordinated by PCA with varimax rotation can be used to select samples in various ways. However, the sample scores on the principal components could be used to constrict the random sampling on some subpopulations of interest. For example, if an experiment requires that the samples should not be very dissimilar in terms of dynamics, one could calculate a percentile (e.g., 25th) cutoff scores and make random sampling only to the subpopulation inside that range. Otherwise, for instance a clustering analysis could provide interesting possibilities for sample selection where one could choose samples which are very similar or dissimilar with respect to this sample space characterization. Also the rank ordering of samples scores combined with equidistant sampling would provide sets of samples where the variations between the samples in each set would represent the ranges of variations in each component. This sample selection method could also be used to perceptually validate the interpretation of principal components as shown by Alluri *et al.* [16]. The perceptual validation of the principal component spaces presented in the current investigation is left for future work but is acknowledged to be an important step in the development of this audio corpus and the proposed stimulus selection framework.

4. DISCUSSION

Anechoic recordings are continuously used in auralizations in room acoustics and related fields. While in many studies the stimuli have been succesfully selected by relying on researchers' expert knowledge and intuition, such practice is susceptiple to experimenter bias and problematic for scientific work. Here, a framework for stimulus production and selection procedure was developed to alleviate this issue, but the applicability of this work remains to be validated in practice. The proposed stimulus production method takes advantage of the possibility to automatically segment (and combine) anechoic instrument recordings of symphonic music. The resulted anechoic audio corpus consists of a large number of short segments of anechoic instrument sounds. The segments contain not only individual notes but also short passages and transitions between notes. Sounds of 15 different instruments of a symphony orchestra are currently included in the corpus which is available online for academic purposes.

Only four short pieces of symphonic music were employed in the production of sound samples, what evidently restricts the representativeness of the audio corpus at the moment. Nevertheless, the proposed framework for a stimulus selection procedure is not restricted to only this sample space but can be implemented in a wide range of studies where the experiment is not bound to a specific type of stimuli. Although the issue of generalizability and representativeness of the test stimuli may still remain even when the proposed approach is advocated, it is already a major advantage to have an explicitly described systematic approach as a stimuli selection procedure, instead of just relying on intuition and subjective opinion. Also the characterization of the stimulus space enables the experimenter to make experimental designs and corresponding statistical models where the influence of the properties of the anechoic stimuli can be analysed. This approach can be also used as a tool to guide the experimenter, even though the final selection is performed on a subjective basis.

In the current work, the proposed framework was used produce two large sets of 5 second long sound samples from four short pieces of anechoic symphonic music. The first set consisted of segments of music where the music was left as composed - albeit cut from the context due to the desired length of stimuli. Clearly, the length of the stimuli used here did not allow for the evaluation of numerous intrinsic and essential aspects of music, such as, longer phrases, verses, choruses, motifs, harmonic developments etc. Such compositional properties of music were not targeted in the current work, where the focus was on lower level acoustic properties in the signals. Extending the length of the extracted segments would be straightforward and would also allow for analysing a number of higher level structures although these structures would be effectively restricted by the musical material. An interesting alternative would be to randomly concatenate the segments of individual instruments to produce random "streams" of instrument sounds, which could be in turn combined with other instrument streams. Such random "music" could be used as a contrastive stimulus to be compared with composed music and for other explorative investigations of higher level musical percepts.

A complementary procedure and a second stimulus set was produced by segmenting the separate instrument recordings and randomly combining these segmented parts. This way also the number of instruments in the produced stimuli could be controlled although this option was not exploited in the current work. Again, the length of the stimuli was fixed to 5 seconds, but variable length

stimuli could also be easily produced. Considering the second method which produces stimuli which are not distinctively musical, calculating features designed to capture perceptually relevant musical features is ambiguous. Other issues which should be addressed in the future work are, for instance, the effect of the window sizes used in calculating the features and determining the most relevant features which would capture the most essential properties of the anechoic signals. The feature set used in the current implementation was limited to 14 features, excluding features, such as, mel-frequency-cepstral coefficient (and its derivatives), pulse clarity, fluctuation centroid and fluctuation entropy. A closer look at these and other features will be taken in the future.

In sophisticated auralization schemes, such as the ones used for studying concert hall acoustics [17], and with appropriate experimental design, this framework could be used to reveal systematic dependencies between the room acoustical properties and the musical features of the source signals. In addition one could also analyze the influence of the instrumentation in the orchestra on the musical features and/or perception of auralization or other signal processing algorithms.

At the moment, the value of the proposed stimuli selection procedure is theoretical and it remains to be studied if in the same experiment the stimuli produced and selected by this procedure will result in a significantly different outcome than the stimuli hand-picked by the experimenter. Researchers are encouraged to explore these and other possibilities as the pre-processed and segmented instrument files and the full length recordings are freely available online. In addition, some basic Matlab scripts for the segmentation and feature extraction are available by request, but potential users are strongly encouraged to write their own scripts as MIRtoolbox is well documented and provided with an extensive user manual.

5. ACKNOWLEDGMENTS

Thanks to Tapio Lokki for discussions about the topic and valuable remarks on this manuscript. Academy of Finland [257099] is thanked for financial support. Also many thanks to the great number of anonymous reviewers!

6. REFERENCES

[1] "University of Ferrara," Available at http://acustica.ing.unife.it/eng-ver/ricerche-eng/Architectural.html, accessed October 30, 2013.

[2] "University of Iowa Electronic Music Studios," Available at http://theremin.music.uiowa.edu/MIS.html, accessed October 30, 2013.

[3] "Vienna Symphonic Library," Available at http://vsl.co.at/, accessed October 30, 2013.

[4] J. Pätynen, V. Pulkki, and T. Lokki, "Anechoic recording system for symphony orchestra," *Acta Acustica united with Acustica*, vol. 94, no. 6, pp. 856–865, 2008.

[5] R. H. Baayen, D. J. Davidson, and D. M. Bates, "Mixed-effects modeling with crossed random effects for subjects and items," *Journal of memory and language*, vol. 59, no. 4, pp. 390–412, 2008.

[6] A. Gelman and J. Hill, *Data analysis using regression and multilevel/hierarchical models*, Cambridge University Press, 2007.

[7] H. Fastl and E. Zwicker, *Psychoacoustics: facts and models*, vol. 22, Springer, 2007.

[8] R. S. Bolia, W. T. Nelson, M. A Ericson, and B. D. Simpson, "A speech corpus for multitalker communications research," *The Journal of the Acoustical Society of America*, vol. 107, pp. 1065–1066, 2000.

[9] M. Cooke, Jon Barker, Stuart Cunningham, and Xu Shao, "An audio-visual corpus for speech perception and automatic speech recognition," *The Journal of the Acoustical Society of America*, vol. 120, pp. 2421, 2006.

[10] T. Eerola and J. K. Vuoskoski, "A review of music and emotion studies: Approaches, emotion models and stimuli," *Music Perception*, vol. 30, no. 3, pp. 307–340, 2013.

[11] M. L. Chanda and D. J. Levitin, "The neurochemistry of music," *Trends in cognitive sciences*, vol. 17, no. 4, pp. 179–193, 2013.

[12] O. Lartillot, P. Toiviainen, and T. Eerola, "A matlab toolbox for music information retrieval," in *Data analysis, machine learning and applications*, pp. 261–268. Springer, 2008.

[13] J. Pätynen, *A virtual symphony orchestra for studies on concert hall acoustics*, Ph.D. thesis, 2011.

[14] M. A. Casey, R. Veltkamp, M. Goto, M. Leman, C. Rhodes, and M. Slaney, "Content-based music information retrieval: Current directions and future challenges," *Proceedings of the IEEE*, vol. 96, no. 4, pp. 668–696, 2008.

[15] G. Tzanetakis and P. Cook, "Marsyas: A framework for audio analysis," *Organised sound*, vol. 4, no. 3, pp. 169–175, 2000.

[16] V. Alluri, P. Toiviainen, I. P. Jääskeläinen, E. Glerean, M. Sams, and E. Brattico, "Large-scale brain networks emerge from dynamic processing of musical timbre, key and rhythm," *Neuroimage*, vol. 59, no. 4, pp. 3677–3689, 2012.

[17] T. Lokki, J. Pätynen, A. Kuusinen, and S. Tervo, "Disentangling preference ratings of concert hall acoustics using subjective sensory profiles," *The Journal of the Acoustical Society of America*, vol. 132, no. 5, pp. 3148–3161, 2012.

COMBINING HIGHER ORDER REFLECTIONS WITH DIFFRACTIONS WITHOUT EXPLOSION OF COMPUTATION TIME: THE SOUND PARTICLE RADIOSITY METHOD

Alexander Pohl,
HafenCity University Hamburg
Hamburg, Germany
alexander.pohl@hcu-hamburg.de

Uwe M. Stephenson,
HafenCity University Hamburg
Hamburg, Germany
post@umstephenson.de

ABSTRACT

The simulation of sound propagation in large rooms and urban environments is mainly performed by geometric simulation methods like ray tracing or the Sound Particle Simulation Method (SPSM). Hence, a severe deficiency is that wave effects are not included, especially if screening or diffraction effects are important. A method to introduce diffraction is the Uncertainty relation Based Diffraction (UBD) model, which has been successfully evaluated recently. To find close edges as sources of diffraction, a subdivision of the room into convex subspaces is performed by virtual walls. However, this causes a recursive split-up of Sound Particles (SPs) at each diffraction event. This effect should be compensated by a re-unification of SPs. Therefore, the Sound Particle Radiosity (SPR) has been found that combines the SPSM with an advantage of the radiosity method: the re-unification of sound energy that uses a discretization of the surface into small patches. Now, SPR has been extended to $3D$ for the first time. To increase the available memory and to decrease the computation time, a parallelization has been implemented for the first time. First results indicate that the discretization of the virtual walls into patches yields additional but tolerable errors in the simulation of diffraction. However, even in $2D$, SPR requires a huge memory. To solve this problem in $3D$ remains a great challenge, even more for more complex rooms. Also a method for a convex subdivision to 3D still has to be found.

1. INTRODUCTION

In room as well as in urban acoustics, where the objects are large compared with wavelengths, geometric-energetic simulation methods are applied, like the image source method [1, 2], the SPSM[3], ray tracing [4, 5], or beam tracing[6]. Today, most beam tracing methods assume pyramidal beams[7, 8]. Naturally, all these neglect wave effects. Nevertheless, in the case of SPs, scattering effects are simulated, but diffraction effects are still hardly simulated in a general way. This is a severe deficiency, especially in very jagged rooms with many obstacles as in urban environments, where sound often reaches receivers solely by diffraction. It is important for auralization purposes, too. There are approaches that add first order diffractions based on the rough approximation of the detour law[9] or more accurate wave theoretical approaches[10]. For small reflection orders, it is efficient to combine beam tracing with diffraction[11]. For high reflection orders, even beam tracing becomes inefficient and the SPSM becomes more efficient - especially if a high number of receivers is used[3]. But even with the SPSM, the number of SPs and, hence, the computation time, explodes due to the necessary recursive split-up of SPs. Avoiding the split-up is possible but finally less effective[12]. Thus, a re-unification of SPs is needed. Stephenson proposed[13] that the acoustic radiosity method[14, 15], as known from computer graphics[6], includes such a reunification effect but is restricted to diffuse reflections. Consequently, the SPSM and the radiosity method are combined to the SPR[13, 16] in order to achieve a method that is cable an arbitrary order of specular reflections, scattering and diffraction without an explosion of the computation time. Meanwhile, the UBD diffraction method[12] has been generalized, thoroughly evaluated and combined with the SPR[17].

However, this paper is focused on the algorithmic problems. It is organized as follows: In Sec. 2, the main features of the SPSM are described, whereas the handling of scattering and diffraction is briefly described. Sec. 3 describes the convex subdivision procedure, which is the base for an effective simulation of diffraction. Sec. 4 describes the SPR before Sec. 5, 6 and 7 describe an estimation the SPR efficiency, a method of parallelization and the accuracy of the sound intensities that are computed with the SPR one after another.

2. SOUND PARTICLE SIMULATION METHOD

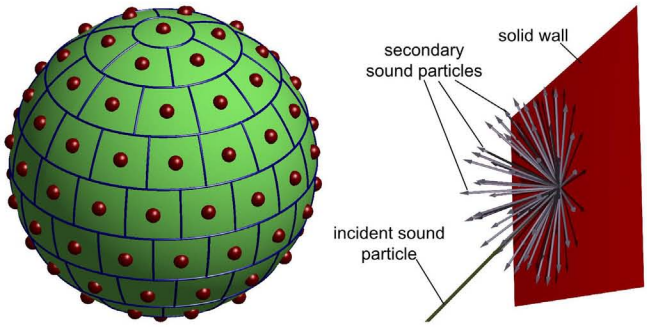
The SPSM is a typical Monte Carlo method, i.e., the idea is to emit a large number of SPs and to trace them iteratively over a number of reflections. To find the next reflection point, each time a number of walls has to be checked for intersection. Furthermore, the point of intersection on the intersected wall has to be determined. Usually, the termination criterion is related to a maximum number of reflections or the desired length (time range) of the echogram. The number of emitted SPs depends on the desired accuracy (e.g., the uncertainty in the computed sound levels) or the spatial resolution of the room surface.

2.1. Simulation of Multiple Frequency Band Simultaneously

In general, SP propagation paths are strongly dependent on the frequency, such that they are computed for each frequency band independently. In the SPSM, all bands are computed simultaneously. Therefore, SPs are carriers of multiple energies instead of carriers of a single sound energy. Thus, the frequency dependent effects like scattering, diffraction, absorption and air attenuation are only allowed to modify the carried energy, but not the actual sound propagation path.

2.2. Emission of Sound Particles

An equal distribution is simple to realize in $2D$, whereas in $3D$ the unit sphere has to be subdivided into (at least approximately) equally sized regions. A good approximation can be found, e.g., by the EQ Sphere method[18] (see Fig. 1(a)).



(a) *primary* (b) *secondary*

Figure 1: *Emission of SPs into about equal solid angle ranges without randomization.*

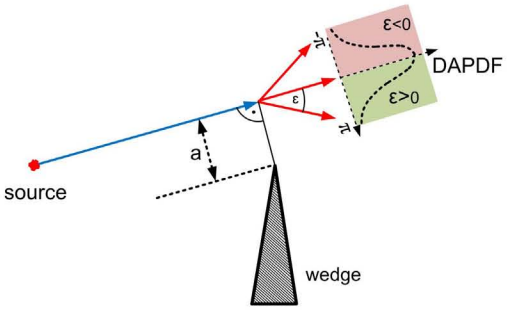


Figure 2: *Diffraction of a SP passing by a single wedge in a distance a by an angle of ε (after Stephenson [12]).*

2.3. Wall Reflection

In case of ideally smooth surfaces, the sound energy is reflected specularly (angle of incidence = angle of reflection). Then, the energy of the SP is reduced by $(1-\alpha)$, where α is the absorption coefficient. In general, α is frequency dependent and, thus, the energies that are carried by the SP have to be modified for each frequency band independently.

2.3.1. Scattering

In case of a rough surface, the ratio of the non-geometrically reflected energy and the total reflected energy is defined as a frequency dependent scattering coefficient σ[19]. To handle scattering in the SPSM, i.e., to define a simple angular characteristic for partially scattering surfaces, commonly used methods[20] interpolate between the geometric reflected energy and the scattered energy distribution according to Lambert's law:

$$\text{in 2D: } \frac{dp}{d\vartheta} = \frac{cos\,(\vartheta)}{2} \text{ and } \quad \text{in 3D: } \frac{dp}{d\Omega} = \frac{cos\,(\vartheta)}{\pi}, \qquad (1)$$

where p is the angular probability density, ϑ the polar and Ω the solid angle, respectively.

In order to avoid a split-up of SPs, these angular probability density functions have been used to compute the direction of the reflected SP. In order to achieve a higher spatial resolution and to keep the sound propagation paths independent of the frequency, a number of S secondary SPs are emitted in equally sized regions (see Fig. 1(b)). The same algorithm as in Sec. 2.2 is used to compute these regions. Each of this SPs carries energies according to an integral over a respective part of the angular probability density function.

2.3.2. Diffraction

In order to introduce edge diffraction into the SPSM, Stephenson proposes to use a diffraction model, which is based on the uncertainty relation[12]. The basic assumptions are: a) edges (for simplification only inner edges) are the main objects where diffraction occurs, b) the SP deflection obeys the uncertainty relation and c) the whole model remains an energetic one. The result is that SPs are diffracted the stronger, the closer (local uncertainty) they pass by an edge (see Fig. 2).

Their deflection is described by a so called Diffraction Angle Probability Density Function (DAPDF), which is derived from the Fraunhofer diffraction at a slit. This UBD diffraction model has been investigated and confirmed in detail during the last years in $2D$, whereas also some approaches exist[17] for the $3D$ case.

This diffraction module is called whenever a virtual wall (see Sec. 3) is intersected by a SP, which indicates one or several close edges. In practice, S secondary SPs that carry energies according to an integral of the DAPDF over the corresponding angle range are emitted. In terms of the algorithm, one difference between scattering and diffraction is that SPs are reflected in the case of scattering, whereas they are transmitted through the virtual wall in the case of diffraction (see Fig. 1(b)). Furthermore, the DAPDF is used instead of Lambert's angular probability density functions.

2.4. Detection of Sound Particles

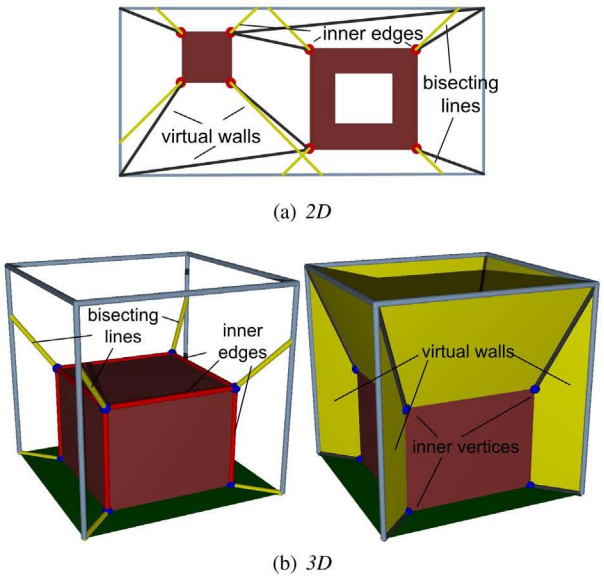
Ideally thin rays never intersect with point-like receivers. So, small detectors around the receivers are created. In contrast to classical ray tracing, the SPSM takes the distance that a SP travels within this detector into account[17]. Thus, the detectors may have any shape. With rectangular detectors, a dense grid of detectors can be established to simulate an immission area[3]. Maps of different room acoustical parameters can be computed on such a grid. These objective quantities are aimed at and analysed here.

3. CONVEX SUB-DIVISION

Most geometrical acoustic simulation methods use a spatial subdivision technique to decrease the computational effort. In the SPSM, a sub-division into convex subspaces is used, which are interconnected by virtual, i.e., acoustically transparent, walls. The main advantages of this approach are that a) the computation time can be reduced by using only convex polyhedra and b) the virtual walls can be used to detect SPs that have to be diffracted[21].

The spatial data structure in $2D$ is given by a **closed** polygon that delimits the sound propagation area. This polygon is constructed from a set of vertices. As the $2D$ scene is interpreted as a cross-section of the $3D$ space, the vertices that protrude into the sound propagation space are called *inner edges*. Although diffraction occurs on all edges, only inner edges are sources of diffraction in the present paper. Thus, these inner edges have to be the starting point for a convex sub-division (see Fig. 3(a)).

Bisecting lines can be defined at the centre direction of each inner edge (see Fig. 3(a)). In order to avoid additional vertices



(a) *2D*

(b) *3D*

Figure 3: *Convex Sub-Division by virtual walls.*

(that would increase the complexity again), these bisecting lines are shifted to the closest vertex of the intersected wall. The resulting bisecting lines act as virtual walls. The polygon is divided into two independent sub-spaces at the virtual wall and the algorithm is repeated until no inner edges remain. This algorithm has been implemented and investigated in detail[21]. The result is that the computation time of the SPSM is almost independent of the number of vertices and only depends on a shape factor that describes the ratio between virtual walls and real walls[21].

In $3D$, the spatial data structure is given by a closed polyhedron that is defined by a set of polygons. Therefore, both inner edges and *inner vertices* can be identified (see Fig. 3(b)). Bisecting lines can be defined on each inner vertex and are shifted to the closest vertex of the intersected wall. Virtual walls can be inserted using these constructed lines, but some crucial questions remain open. Here, the result is a set of convex polyhedra that are connect by virtual walls. As this algorithm is not fully implemented in $3D$ yet, both a manual sub-division as well as an automatic sub-division into tetrahedra[22] are used preliminarily. Again, the computation time is dependent on a similar shape factor, describing the ratio of the virtual wall surface and the real surface.

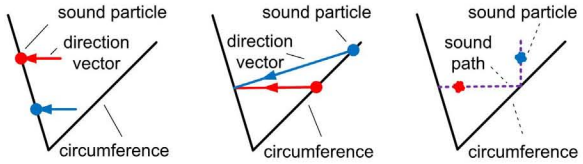
4. SOUND PARTICLE RADIOSITY METHOD

Any recursive split-up of SPs causes an exponential increase of their number and, thus, the computation time. This can only be compensated by a reunification of the SPs. The idea of the radiosity method is to consider a radiation exchange between pairs of small patches of the room surface ending up in solving (directly or iteratively) just a linear equation system[15].

Meanwhile, the radiosity method has been applied and improved in room acoustics. It is either combined with beam tracing to compute the late (diffuse) reflections in a hybrid method[14] or appropriate discretizations without the extension to specular reflections are investigated[23]. Another hybrid method combines the radiosity method with the image source method to allow specular reflections of first order[24]. Furthermore, a generalization of the energy exchange in the radiosity method to specular reflections is presented[25]. However, this method extends the radiosity method by specular reflections on the contrary to an aspired equal weighting of specular and diffuse reflections of the SPR. In order to introduce reunification into the SPSM, Stephenson used the feature of the radiosity method that the number of sound propagation paths, i.e., the number of energy exchange factors, is finite[13]. In the radiosity method, this is achieved by neglecting the history of the sound energy. Stephenson interpreted that as reunification of sound energy on a patch. However, this is the reason that only totally diffuse reflections are possible - the drawback of the radiosity method. In order to allow a generalization to specular reflections, the sound energy is reunified on sound paths **between pairs of patches** rather than on single patches[17] in case of the SPR. Thus, the angle of incidence is not lost anymore. For Stephenson, this is a discretization of both the surface (radiosity method) and the directional space (SPSM)[13]. As a result, the SPR allows the simulation of specular reflections, scattering and even diffraction without an explosion of the computation time.

4.1. Discretization

Three parameters of a SP have to be discretized (see Fig. 4).



(a) *target point* (b) *emission point (direction of incidence)* (c) *impact time*

Figure 4: *Three parameters describe the SP propagation path.*

To discretize the first two parameters, the circumference of the room is divided into small patches. If SPs intersect with a patch, their intersection points are shifted to the centre of the patch. In addition, the time of incidence on a patch is discretized in time intervals Δt, whose respective traveling distance $c \cdot \Delta t$ is chosen to equal the length of a patch. As SPs that travel less than a time interval might cause infinite loops, these travelled distance are increased to one time interval. As a result, a discretized SP propagation path is completely described by three numbers. A SP only carries energy (or a group of energies for different frequency bands).

4.2. Sound Particle Logistics

In this method, the reunification of SPs happens when they intersect with the room surface. This is plausible, because this is the time when they change their direction or split-up. To allow reunification (in contrast to the SPSM), the SPs have to be traced quasi simultaneously. The problem is: From an algorithmical point of view, this is not possible exactly. Before SPs are traced further from their intermediate position on a patch, they must *wait*, until all *older* SPs on the way have reached this position. In other words, all SPs have to travel a certain distance, before the first SP is allowed to travel further. This processing order can only be

achieved by a completely new frame algorithm. This is performed by a Reunification Matrix (RUM).

4.3. Reunification Matrix (RUM)

The RUM temporary stores SPs and extracts them when needed. It is a multi-dimensional storage of energy, where the parameters of the SPs (the number of the starting patch, the number of the end patch and the number of the impact time interval) are encoded in the position of the energy (see Fig. 5)[16].

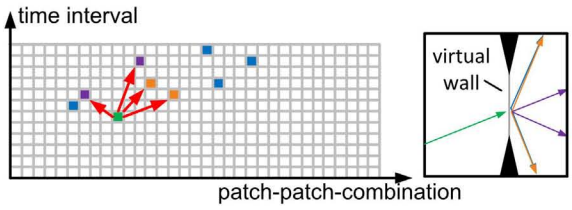


Figure 5: *Example for the SPR process in Reunification Matrix (RUM). To achieve a 2D-depiction, the number of* ***combinations of the starting point and the end point*** *is shown as the $x-axis$ and the number of the time interval is shown on the $y-axis$.*

Initially, the RUM elements that describe the emission from the sound source are filled with energy (normalized to the sum of 1). Later, always the *oldest*, i.e., least far travelled SP energy (*green element*), is taken out of the RUM. The sound propagation path is reconstructed by the position of the energy. This particle is processed by applying the laws of reflection, diffraction and scattering known from the SPSM. After each step, the SP energy is stored in the RUM at the place corresponding to the intersected patch, the former intersected patch and the time interval of incidence (*red arrows*). Whenever energy is to store in an already occupied element (*orange elements*). Thus, SPs are reunified.

This procedure is repeated, until all energies are transported to the uppermost RUM elements according to the maximum travelling time.

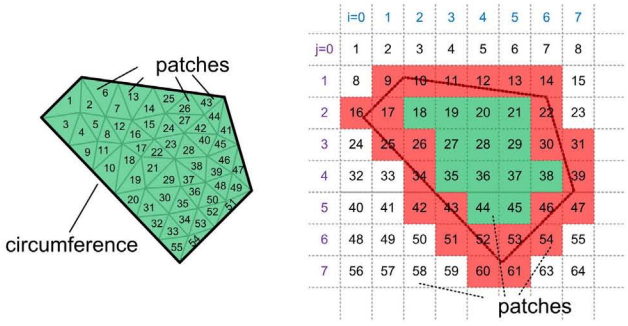
4.4. Definition of Patches

In $2D$, the patches are equally sized line segments of length l_P. In order to have a discretization parameter that is independent of the room size, a relative parameter f_P is introduced

$$f_P = \frac{l_P}{\bar{l}}, \tag{2}$$

where $\bar{l}$ is the mean free patch length of the room. However, in $3D$, a surface instead of a line has to be discretized into patches. One attempt is to perform a triangulation of the wall into equally sized triangles by a refined Delaunay triangulation[26] (see Fig. 6(a)). The drawback is that identifying single triangles and computing their patch number is time consuming.

A huge simplification and more efficient is to place a rectangular grid on the surface (see Fig. 6(b)). Only the determination of the patch centres of patches that contain the wall boundary is demanding. The parameter f_P describes the average patch length.



(a) *Delaunay triangulation (after [26])* (b) *regular grid*

Figure 6: *Different techniques defining patches on a surface.*

5. EFFICIENCY OF THE SOUND PARTICLE RADIOSITY ALGORITHM

To determine the speed-up of the SPR relative to the SPSM, two contrary conditions have to be taken into account. On the one hand, the storage of SPs in the RUM causes additional computational effort, whereas, on the other hand, the reunification of SPs reduces the computational effort. Thus, the reunification rate is the main parameter describing the efficiency of the SPR. Furthermore, the computation time of the SPR is investigated to determine the computational effort of reunification.

5.1. Reunification Rate

First, the number of occupied elements $N_{RUM}\,(o)$ after a reflection order o is determined. In this context, reflection order means that each SP is reflected (including scattering or diffraction) o times.

On average, every SP is split-up into $1+S$ secondary SPs when they intersect with a wall. So, the number of occupied matrix elements reads **without reunification**

$$N_{RUM}\,(o) = N \cdot (1+S)^o\,, \tag{3}$$

where N is the number of emitted SPs. With reunification[16], the number of occupied matrix elements is decreased. A recursive formulation reads

$$\begin{aligned} N_{RUM}\,(o+1) &= N_{RUM}\,(o) \cdot q^{N_{RUM}(o)} \\ &+ S \cdot \frac{1-q^{N_{RUM}(o)}}{1-q} \\ N_{RUM}\,(o=0) &= N \text{ with } q = 1 - \frac{S+1}{K_{RUM}}, \end{aligned} \tag{4}$$

where K_{RUM} is the available number of elements in the RUM.

A comparison is given in Fig. 7 for different S.

Without reunification, the number of SPs increases exponentially with the number of computed reflections o. With reunification, the number of occupied matrix elements and, thus, the number of simultaneously existing SPs, is reduced significantly. $N_{RUM}\,(o) = K_{RUM}$ acts as the upper limit of the occupied matrix elements and, hence, SPs.

The size of the RUM is proportional to the product of a) the number of starting patches, b) the (same) number of end patches

Figure 7: *Number of occupied elements $N_{RUM}\,(o)$ as a function of the reflection order o, with different split-up values S. The estimated number of occupied elements with reunification (solid, SPR) is compared to the case without reunification (dotted, SPSM)*

and c) the number of time intervals. The number of time intervals is reduced by recycling the RUM after a maximum free path length (room diagonal) has been travelled. For a quadratic (cubical) room, the number of available matrix elements turns out to be

$$K_{RUM,2D} \approx 47 \cdot \frac{1}{f_P^3} \text{ and } K_{RUM,3D} \approx 474 \cdot \frac{1}{f_P^5} \qquad (5)$$

In case of a more complex scene, this size increases only linearly with the number of convex sub-spaces (another advantage of the sub-division). A modern computer with $16GB$ RAM allows maximum discretization factors of $f_P \approx \frac{1}{450}$ in $2D$ or $f_P \approx \frac{1}{25}$ in $3D$.

5.2. Computation Time

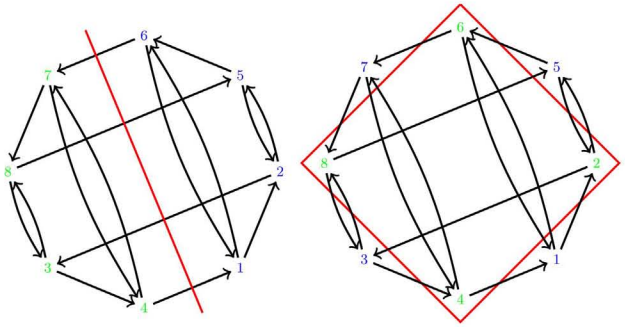
The computation times have been measured by a simulation of a two dimensional, rectangular room with $N = 1000$ primary SPs and a split-up of $S = 25$. The computation times of the SPR and the SPSM are compared in Fig. 8.

Figure 8: *Computation times of Sound Particle Radiosity (SPR).*

As expected, the computation time of the SPSM increases exponentially with the reflection order (note the logarithmic $y - scale$). Compared to this, the computation time of the SPR is reduced. With this set-up, both computation times are equal at approximately $o_{Max} = 4$. For higher reflection orders, the SPR is faster than the SPSM, but the SPSM is still faster for lower reflection orders. For only one reflection, the SPSM computes less than a second, whereas SPR already has a computation time of approximately one minute. The reason for this behaviour is the additional computational effort to access the RUM.

6. PARALLELIZATION

Besides the reduction of computation time, another main purpose of parallelization is to add additional memory by using computer clusters[27]. For efficiency, the idea is to assign a part of the RUM to every computer. A simple approach is to assign all RUM elements that describe the sound propagation paths within the same convex sub-space to the same computer. The bottleneck of these computer clusters is the communication between distributed computers. To reduce this, a good decomposition of the RUM is needed. In the RUM, each *column* (see Fig. 5) represents a sound propagation path, i.e., a patch-patch combination. The nodes in the graphs (see Fig. 9) depict RUM elements and the arrows sequences of sound propagation paths, i.e., communications between matrix elements. The boundary between RUM elements of different computers (in reality: maybe of different sub-spaces) is indicated by a *red* edge (see Fig. 9). Every transfer between different (sub-)RUMs is called edgecut. The number of these edgecuts indicates the amount of communication overhead (remote calls), which has to be reduced. For an example of eight *columns*, two different decompositions are shown in Fig. 9.



(a) 4 *edgecuts* (b) 16 *edgecuts*

Figure 9: *Two decompositions of eight matrix elements in a graph model.*

A good decomposition is one that minimizes the remote calls. In our experiments, a standard toolbox has been used[27]. For further experiments, however, it is advisable to exploit geometrical information, i.e., the transfer probabilities between pairs of patches. They are increased for patch-patch-combinations within parallel and specularly reflecting walls (causing flutter echoes) or reduced in case of patch-patch-combinations that are distributed among distant sub-spaces.

7. ACCURACY OF SOME COMPUTED ROOM ACOUSTICAL PARAMETERS

Due to the discretization of the sound propagation paths, numerical errors are produced - especially within the simulation of diffraction. In order to determine the accuray of the SPR method, three parameters are computed: a) the relative error of the total sound intensity, b) the relative error of a short time intervall of an echogram and c) the relative error of the reverberation time. The result of a SPSM simulation, i.e., without discretization, serves as reference. All errors are strongly dependent on the discretization parameter f_P.

7.1. Influence of the Patch Size in 2D

In a first attempt, simulations have been performed without scattering or diffraction to focus on the numerical error due to the discretization.

The geometrical scene is defined by a rectangular room with walls of length $a = 10m$ and 81 receivers with a diameter of $r_D = 1m$. The maximum simulation time, and thus the maximum length of the echogram, is restricted to $T_{Max} = 0.1s$ to focus on single reflections. The echogram is split up into 1000 time intervals of $\Delta t = 0.1ms$. The absorption degrees are set to $\alpha = 0.5$ and scattering is disabled ($\sigma = 0.0$, $S = 0$) on all four walls. The echogram that is simulated with the SPR converges to the echogram that is simulated with SPSM for decreasing discretization parameters f_P. Only slight differences occur for $f_P = 1/100$, but strong deviation occur for $f_P = 1/20$.

For all receivers, the reverberation time and the total sound intensity are computed with the SPR are their errors are defined relative to the results of the SPSM. Both values are below 2% for $f_P < 1/100$ and increase only up to 20% for $f_P = 1/10$. The reason is that small variations of the sound propagation path only slightly affect both the reverberation time and the total intensity. However, the average error in a single time slot is 300% for $f_P = 1/20$ and 50% for $f_P = 1/200$. Here, slight time shifts cause a detection in adjacent time intervals. (A graph is omitted, because the echograms are similar to Fig. 10).

7.2. Influence of the Patch Size in 3D

The same experiment has recently been performed for a cubical room. The dimensions are adjusted to $a = 11.78m$ and $r_D = 1.178m$ in order to have the same mean free path length as in the former experiment. An example of one of the 729 receivers (filling the whole cube) is shown in Fig. 10. The result is similar to the

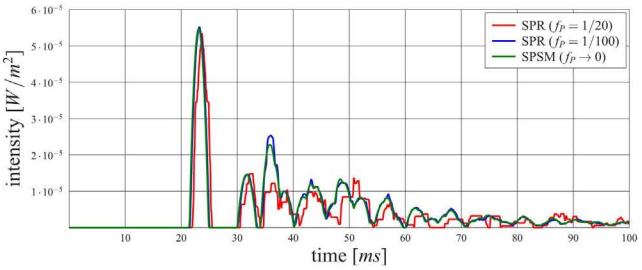
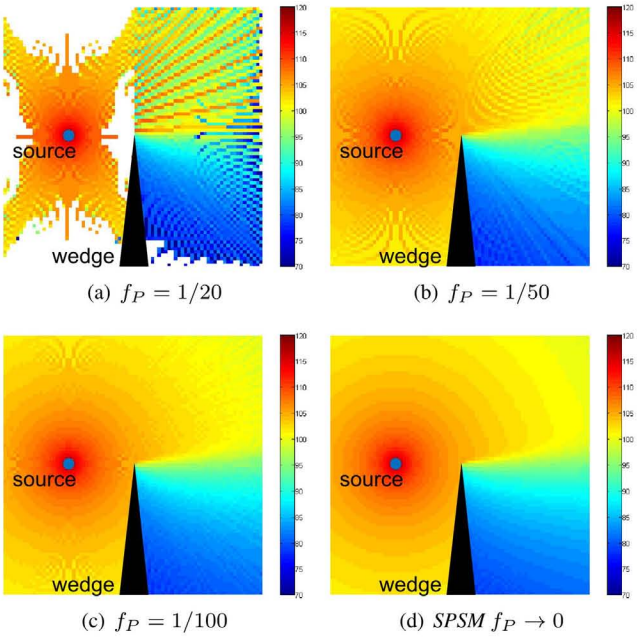


Figure 10: *Echograms of different discretizations in 3D.*

$2D$ result, which is confirmed by the numerical results of the error in reverberation time and total sound intensity. Quantitatively, the average error in a time interval in increased by a factor of 2 compared with the $2D$ result.

7.3. Influence of the Discretization on the Simulation of Diffraction

In case of diffraction, the spatial behaviour of the numerical errors is more important than its time behaviour. Due to the functional principle of SPR, all SPs are discretized by shifting their intersection points to the patch centres before they are diffracted or detected. Former investigations revealed that at least one SP has to be diffracted in the region of $d = 0.1\lambda$ above the wedge. This region is in the range of $0.002m(16.000Hz) < d < 1.092m(31.5Hz)$ (Here: $6.8cm$ at $500Hz$). The simulations have been performed in a $2D$ rectangular room with a width of $a = 20m$ in each dimension. A single wedge with a height of $a/2 = 10m$ is placed on the centre of the floor. The source is placed in the centre of the subspace on the left-hand side of the wedge (see Fig. 11). No specified receiver position is needed, because a sound intensity map is computed for a whole receiver grid. These receivers are placed is a distance of $w_{grid} = 0.1m$ that equals their diameter ($200x200$ receivers). All surfaces (incl. wedge) are fully absorbent ($\alpha = 1.0$) to focus on the effect of diffraction). A sufficient number of $N = 500.000$ primary SPs are emitted and split up into $S = 200$ SPs at each diffraction[17].



(a) $f_P = 1/20$ (b) $f_P = 1/50$

(c) $f_P = 1/100$ (d) *SPSM* $f_P \to 0$

Figure 11: *Sound intensity levels for different discretizations of a diffraction simulation at $f = 500Hz$ in a quadratic, fully absorbent room.*

Again, the numerical error decreases with decreasing discretization parameters. To describe the influence of the discretization error on diffraction quantitatively, the average over the absolute level difference is investigated for all frequency bands (relative to the respective SPSM) in the shadow zone ($1dB$ at $f = 500Hz$ and $f_P = 1/50$).

The error is shown to be strongly influenced by the ratio of the patch length and the wavelength and only hardly dependent on the absolute patch size. As a result, the patch length has to be adjusted to meet the requirements for the highest frequency simulated. For an accuracy of $1dB$ even at $4kHz$, a discretization parameter of $f_P \leq \frac{1}{250}$ is required.

8. CONCLUSION AND OUTLOOK

Due to the reunification effect, the SPR allows a room acoustical simulation including diffraction or scattering of an arbitrary order without an explosion of the computation time. It has been implemented in $3D$ and on a computer cluster for the first time. The SPR

frame algorithm is identical in $2D$ and $3D$. The spatial discretization of the surface has been performed using quadratic patches for the first time. However, the size of the RUM increases to the power of 5 in $3D$ instead of the power of 3 in $2D$ with the discretization parameter. This huge size is still the main drawback of the SPR, which can be compensated only by part using parallelization.

The concept of the convex sub-division still has to be fully implemented in $3D$. The diffraction module is already analytically defined and tested in $3D$ for simple setups[17] and a combination with the SPR algorithm is in preparation. However, the main challenge is to find a method to reduce the required memory.

9. REFERENCES

[1] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *The Journal of the Acoustical Society of America*, vol. 65, no. 4, pp. 943–950, 1979.

[2] J. Borish, "Extension of the image model to arbitrary polyhedra," *The Journal of the Acoustical Society of America*, vol. 75, no. 6, pp. 1827–1836, 1984.

[3] Uwe M. Stephenson, "Comparison of the mirror image sourc emethod and the sound particle simulation method," *Applied Acoustics*, vol. 29, no. 1, pp. 35–72, 1990.

[4] A. Krokstad, S. Strom, and S. SÃ¸rsdal, "Calculating the acoustical room response by the use of a ray tracing technique," *Journal of Sound and Vibration*, vol. 8, no. 1, pp. 118 – 125, 1968.

[5] Michael Vorländer, "Simulation of transient and steady-state sound propagation in rooms using a new combined ray-tracing/image-source algorithm," *The Journal of the Acoustical Society of America*, vol. 86, no. 7, pp. 172–178, 1989.

[6] Andrew S. Glassner, Ed., *An introduction to ray tracing*, Academic Press Ltd., London, UK, 1989.

[7] A. Farina, "RAMSETE - a new Pyramid Tracer for medium and large scale acoustic problems," in *Proceedings of the Euronoise*, Lyon, France, 1995.

[8] Thomas Funkhouser, Ingrid Carlbom, Gary Elko, Gopal Pingali, Mohan Sondhi, and Jim West, "A beam tracing approach to acoustic modeling for interactive virtual environments," in *Proceedings of the 25th Annual Conference on Computer Graphics and Interactive Techniques*, New York, NY, USA, 1998, SIGGRAPH '98, pp. 21–32, ACM.

[9] Z. Maekawa, "Noise reduction by screens," *Applied Acoustics*, vol. 1, no. 3, pp. 157–173, 1968.

[10] U. P. Svensson, R. I. Fred, and J. Vanderkooy, "An analytic secondary source model of edge diffraction impulse responses," *The Journal of the Acoustical Society of America*, vol. 106, pp. 2331–2344, 1999.

[11] Nicolas Tsingos, Thomas Funkhouser, Addy Ngan, and Ingrid Carlbom, "Modeling acoustics in virtual environments using the uniform theory of diffraction," in *Proceedings of the 28th annual conference on Computer graphics and interactive techniques*, New York, NY, USA, 2001, SIGGRAPH '01, pp. 545–552, ACM.

[12] Uwe M. Stephenson, "An energetic approach for the simulation of diffraction within ray tracing based on the uncertainty relation," *Acta Acustica united with Acustica*, vol. 96, pp. 516–535, 2010.

[13] Uwe M. Stephenson, "Quantized pyramidal beamtracing or a sound-particle-radiosity- algorithm? - new solutions for the simulation of diffraction without explosion of computation time," in *Proceedings of the Research Symposium on Acoustic Characteristics of Surfaces*, Salford, United Kingdom, 2003.

[14] T. Lewers, "A combined beam tracing and radiatn exchange computer model of room acoustics," *Applied Acoustics*, vol. 38, no. 2–4, pp. 161–178, 1993.

[15] H. Kuttruff, "A simple iteration scheme for the computation of decay constants in enclosures with diffusely reflecting boundaries," *The Journal of the Acoustical Society of America*, vol. 98, no. 1, pp. 288–293, 1995.

[16] Alexander Pohl and Uwe M. Stephenson, "A combination of the sound particle simulation method and the radiosity method," *The Journal of Building Acoustics*, vol. 18, no. 1, pp. 97–122, 2011.

[17] Alexander Pohl, *Simulation of Diffraction Based on the Uncertainty Relation – An Efficient Simulation Method Combining Higher Order Diffractions and Reflections*, Ph.D. thesis, HafenCity University Hamburg, 2014.

[18] Paul Leopardi, *Distributing points on the sphere: Partitions, separation, quadrature and energy*, Ph.D. thesis, The University of New South Wales, 2007.

[19] "ISO 17497 Part 1: Measurement of random-incidence scattering coefficient in a reverberation room," 2004.

[20] Dirk Schröder and Alexander Pohl, "Modeling (non-) uniform scattering distributions in geometrical acoustics," in *Proceedings of International Congress on Acoustics*, Montreal, Canada, 2013.

[21] Alexander Pohl and Uwe M. Stephenson, "Efficient simulation of sound propagation including multiple diffractions in urban geometries by convex sub-division," in *Proceedings of Internoise*, Lisbon, Portugal, 2010.

[22] Hang Si, *Three Dimensional Boundary Conforming Delaunay Mesh Generation*, Ph.D. thesis, Technische Universität Berlin, 2008.

[23] Eva-Marie Nosal, Murray Hodgson, and Ian Ashdown, "Improved algorithms and methods for room sound-field prediction by acoustical radiosity in arbitrary polyhedral rooms," *The Journal of the Acoustical Society of America*, vol. 116, no. 2, pp. 970–980, 2004.

[24] N. Korany, J. Blauert, and O. Abdel Alim, "Acoustic simulation of rooms with boundaries of partially specular reflectivity," *Applied Acoustics*, vol. 62, no. 7, pp. 875 – 887, 2001.

[25] Samuel Siltanen, Tapio Lokki, Sami Kiminki, and Lauri Savioja, "The room acoustic rendering equation," *The Journal of the Acoustical Society of America*, vol. 122, no. 3, pp. 1624—1645, 2007.

[26] Jonathan Richard Shewchuk, "Robust Adaptive Floating-Point Geometric Predicates," in *Proceedings of the Twelfth Annual Symposium on Computational Geometry*. 1996, pp. 141–150, Association for Computing Machinery.

[27] Alexander Pohl, Jan Winkelmann, and Uwe M. Stephenson, "Parallel sound particle radiosity: Reunification of diffracted and scattered sound particles on parallel computers," in *Proceedings of DAGA*, Meran, Italy, 2013.

Proc. of the EAA Joint Symposium on Auralization and Ambisonics, Berlin, Germany, 3-5 April 2014

ACHIEVING REALISTIC AURALISATIONS USING AN EFFICIENT HYBRID 2D MULTI-PLANE FDTD ACOUSTIC MODEL

Stephen Oxnard, *

University of York
Audio Lab, Department of Electronics
York, UK
so523@york.ac.uk

Damian Murphy,

University of York
Audio Lab, Department of Electronics
York, UK
damian.murphy@york.ac.uk

ABSTRACT

This research examines the validity of utilising a 2D multiplane FDTD acoustic model to simulate low frequency sound propagation as part of a hybrid room impulse response (RIR) synthesis system. Analytic results, pertaining to the comparison of simulated low frequency multiplane RIRs with both practical RIR measurements and 3D FDTD simulated RIRs, demonstrate that a good level of accuracy is attained through use of this hybrid modelling paradigm. This claim is further supported, in part, by comparative subjective test results. Furthermore, 2D multiplane simulations are shown to be far more efficient than full 3D FDTD modelling procedures as they achieve a ∼98% reduction in computation time.

1. INTRODUCTION

Traditionally, geometric acoustic modelling approaches have been harnessed to virtually simulate sonic environments for the purposes of room acoustics prediction and analysis. These approaches, which include ray-tracing, image source method (ISM) and acoustic radiosity (AR), while computationally efficient, produce an inaccurate representation of sound propagation at low frequencies. This arises due to underlying assumptions, common to the implementation of all geometric modelling techniques, which do not facilitate the preservation of wave phenomena such as interference effects, diffraction and standing waves. Conversely, modern wave-based approaches, such as the Finite Difference Time Domain (FDTD) and Digital Waveguide Mesh (DWM) paradigms, allow for direct numerical solution of the wave equation. As such, wave-based acoustic models inherently emulate the behaviour of low frequency sound propagation to a far greater level of accuracy than their geometric predecessors. However, the implementation of these numerical acoustic models is hindered greatly by their reliance on extensive computational resources and lengthy compute times.

Despite the computational challenges posed by numerical approaches, which are particularly apparent for 3D acoustic simulations, implementations of 3D FDTD models may be processed in real-time under certain conditions. A previous study [1] demonstrates that it is possible to render 3D soundfields by means of the FDTD paradigm at interactive sampling rates through utilisation of Graphics Processing Units (GPUs). However, this outcome is realisable only when restricting the size of the modelled spatial domain and/or the simulated frequency bandwidth. An alternative, efficient wave-based modelling method devised by Raghuvanshi et al. [2, 3] applies Adaptive Rectangular Decomposition (ARD) to segment the spatial domain into cuboid sections for which the analytical solution of the wave equation is known. As such, the entire modelled soundfield may be resolved temporally through a series of weighted Cosine basis functions. In [3], the ARD system is shown to achieve significant savings for band-limited acoustic modelling in terms of memory consumption and requires up to 18x less computation compared to FDTD schemes, noting that a GPU implementation was not used. Further reductions in computational expense are attained when reducing the spatial dimensionality of the model. This was discussed in relation to room acoustic modelling by Kelloniemi et al. in [4] where multiple 2D DWMs were used to emulate acoustic simulations of a geometrically simplistic 3D soundfield. In this work, the authors predicted reductions in memory consumption and processing cost of 97% and 99% respectively against a 3D DWM model while preserving important spectral features present in the soundfield.

All numerical acoustic modelling implementations are limited in their use in terms of valid bandwidth due to inherent dispersion error which becomes evident with increasing frequency [5]. For this reason, full bandwidth wave-based models rely on greatly over-sampled spatial domains leading to large increases in computational cost. In attempt to alleviate the trade-off between accuracy and efficiency in full audio bandwidth room acoustic simulations, several examples of hybrid modelling systems have been developed. These methods seek to render room impulse responses (RIRs) through complementary assimilation of two or more virtual modelling paradigms. An example of such a system, documented in [6] combined use of a wave-based 3D FDTD scheme with optimised ISM and AR models for hybrid RIR generation. This modelling approach limited the application of FDTD simulations to low frequencies and, hence, reduced the required computation load and run-times comparative to full audio bandwidth FDTD schemes. Results obtained were then amalgamated with high frequency RIRs generated by the ISM and AR models to render a spectrally complete hybrid RIR. A more recent study [7], largely influenced by the research of [4], sought to validate the use of multiple 2D cross-sectional FDTD schemes as a means of representing low frequency sound propagation throughout a 3D enclosure, again as part of a hybrid modelling approach. Results presented in this work demonstrated that the wave-based 2D multiplane approach achieved a reasonable approximation to low frequency RIRs simulated by a full 3D FDTD scheme, in a simplistic modelling scenario, while reducing simulation run-times by 99.15%.

The research documented in this paper seeks to further examine the validity of utilising 2D multiplane FDTD schemes as an

* Funded by a U. of York EPSRC Doctoral Training Grant.

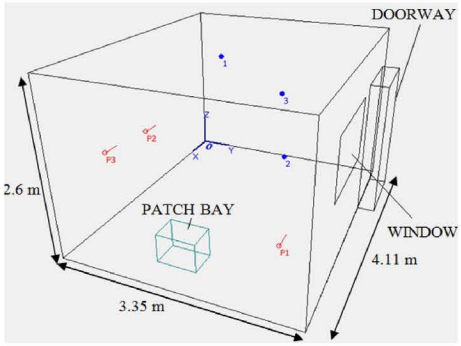


Figure 1: 3D depiction of the Live Room geometry (as viewed in ODEON) highlighting significant dimensions and features and the orientation with the x,y and z axes. Source and receiver placements are represented by 'P1-P3' and '1-3' respectively.

efficient means of low frequency acoustic modelling. This task was undertaken by utilising the multiplane approach, in conjunction with ray-based geometric techniques, to render hybrid RIRs of an existing space. Resulting low frequency RIR spectra were then compared against those of practical RIR measurements and those generated by a 3D FDTD model of the enclosure in order to analyse the level of agreement between both virtual models and reality. Finally, preliminary subjective tests were conducted to compare auralisations rendered by means of 2D multiplane and 3D FDTD hybrid acoustic modelling in a perceptual sense. It is proposed that 2D multiplane FDTD modelling will offer a highly efficient means of augmenting geometric approaches to rendering large-scale acoustic scenes for virtual reality and acoustic prediction applications where perceptually valid, real-time or interactive design workflows are required.

2. THE VIRTUAL ACOUSTIC MODELS

The acoustic models created for the purposes of this study were virtual representations of the recording studio live room situated in the Audio Lab at the University of York. An overview of the room geometry is depicted in Figure 1. Practical RIR measurements were obtained from this space prior to the modelling procedure in order to gain real results for comparison. For the purposes of RIR measurement, an omni-directional sound source was approximated by rotating a Genelec 8130A loudspeaker around the azimuth at increments of 90°and capturing IRs for each orientation using a ST450 Soundfield microphone. Mono RIRs were rendered by summing the W-channel of the captured B-Format responses obtained for each loudspeaker orientation. In this way, three impulse responses were collected from the live room using the sound source and receiver placements detailed in Table 1. In total, three acoustic models, described here, were constructed to simulate the acoustic of this space for each case of source and receiver location, yielding three RIRs per model.

2.1. Geometric Model

ODEON 10.1 Auditorium [8] acoustic prediction software was utilised to develop the geometric acoustic model of the live room and render the mid-high frequency RIRs. This software package is an industry standard acoustic modelling program which combines

CASE	Source (x,y,z) (m)	Receiver (x,y,z) (m)
1	(3.61, 2.85, 0.68)	(0.60, 0.69, 2.00)
2	(2.91, 0.65, 1.49)	(2.91, 2.65, 1.49)
3	(3.61, 0.50, 1.45)	(0.70, 1.85, 1.50)

Table 1: Overview of source and receiver placements in each RIR measurment case.

ray-tracing and ISM to render RIRs for virtual environments developed by the user.

The geometries of the live room were compiled from architectural diagrams and input to ODEON via a .par (parameter) file. Sound source and receiver locations were defined within the model with reference to the locations utilised for practical measurements. Reference receivers were also defined at a distance of 1m from each sound source location for the purposes of the RIR calibration process described in section 3. A total of 7 different material types were applied to the surfaces incorporated in the model: Floor (carpet); Ceiling (fiber board); Walls (plaster board and cavity); Window (double glazed); Door (solid wood); Patch Bay outer shell (medium density wood); Patch Bay front panel (metal). The closest approximate material properties available in the ODEON material library were applied as appropriate. To generate results of suitable accuracy, 50000 rays were used to render each impulse response. Mono omni-directional RIRs were obtained by extracting the W-channel of B-Format impulse responses in each case in order to maintain consistency both with practical measurements and with the capture method implemented in the FDTD schemes.

2.2. 3D FDTD Model

The derivation of the 3D FDTD scheme utilised in this work begins with the 2^{nd} order homogeneous wave equation:

$$\frac{\partial^2 p(\overrightarrow{n},t)}{\partial t^2} = c^2 \nabla^2 p(\overrightarrow{n},t) \tag{1}$$

where p is a measure of acoustic pressure at time t at a location given by the 3D Cartesian positional vector $\overrightarrow{n}$, c is wave speed in ms^{-1} and ∇^2 is the 3D Laplacian operator. Spatio-temporal discretisation of (1) through use of centered finite difference approximations yields the 'Standard Rectilinear' (SRL) update equation, see e.g. [9], implemented in this study. To ensure the FDTD scheme operates with numerical stability a lower limit is imposed on the magnitude of the spatial sampling interval. For 3D FDTD schemes this lower limit h_{3D}, calculated by means of Von Neumann analysis [10], is given as $h_{3D} \geq ck\sqrt{3}$ where k is the discrete time step (s) defined as the reciprocal of the temporal sampling frequency Fs. The spatial discretisation of (1) inherently gives rise to dispersion error, due to anisotropic wave propagation, which is most apparent at high frequencies. In order to reduce the impact of dispersion effects, h_{3D} was set equal to the lower limit giving a usable simulation bandwidth of 0.196Fs [9]. Frequency independent locally reactive surface (LRS) boundary conditions [9] were utilised to terminate the spatial domain allowing for appropriate reflection coefficients to be applied to each bounding surface. Suitable reflection coefficients for each modelled boundary surface were derived by averaging the low frequency absorption coefficients, provided in ODEON, for the corresponding surface material type applied in the geometric model. Table 2 provides the reflection coefficients applied to each surface in the 3D FDTD

Surface	R_{3D}	$R_{2D,MP}$
Floor	0.8860	0.7982
Walls	0.9592	0.9061
Ceiling	0.8944	0.8541
Window	0.9747	0.9442
Door	0.9539	0.9033
Patch Bay Shell	0.9487	0.9002
Patch Bay Front	0.7416	0.6742

Table 2: Overview of reflection coefficients applied to different surface types in the 3D (R_{3D}) and 2D multiplane ($R_{2D,MP}$) FDTD models .

model.

The 3D FDTD model may be envisaged as a rectilinear lattice of pressure sampling nodes occupying the volume defined by the dimensions of the live room with an internodal distance of 0.0135m corresponding to Fs = 44.1kHz. The positioning of boundary surfaces and source and receiver placements within the model was calculated by rounding the actual position to the nearest sampling instance with a maximum deviation of $((3h_{3D})^2/4)^{0.5} < 0.012$m. A Dirac Delta sound excitation signal was applied for each RIR case by initialising the source node to a unity value. Impulse responses were then captured by recording the response to this excitation at the receiver node. Approximately 17 000 000 sampling nodes were required to render the 3D model and average computation times were in the region of 3800s for 1s RIRs at the audio sampling rate.

2.3. 2D Multiplane FDTD Model

The 2D Multiplane FDTD model was designed to approximate the 3D live room geometry through use of three intersecting 2D SRL FDTD schemes derived in an analogous manner to the 3D scheme with $\overrightarrow{n}$ and ∇^2 of (1) reduced to 2 dimensions. Again, a lower limit was imposed on the spatial sampling interval in the interest of maintaining numerical stability, which for the 2D FDTD case is: $h_{2D} = ck\sqrt{2}$. This gives an inter-nodal distance of 0.011m for Fs = 44.1kHz yielding a maximum spatial positioning error of source, receiver and bounding surface placement of $((h_{2D})^2/2)^{0.5} < 0.008$m comparative to practical dimensions. Locally reacting surface boundary conditions were also utilised in the multiplane model, however the reflection coefficients applied to each bounding surface (see Table 2) were subject to calibration as discussed in section 3.2.

Each 2D FDTD scheme represented a cross-section of the space orientated in the x-y, x-z and y-z planes with a common point of intersection defined by the position of the receiver in each measurement case. Sound excitation positions were defined by taking the sound source locations and projecting them perpendicularly onto each plane. In this way, the excitation nodes defined in the x-y, x-z, and y-z orientated cross-sectional schemes shared the (x,y), (x,z) and (y,z) coordinates of the measurement source locations respectively. As with the 3D case, excitation of each plane was achieved by initialising each source node to a unity pressure value. The individual responses captured from the cross-sectional planes were aligned in time, synchronising the direct sound component of each 2D RIR, and then summed to obtain the complete multiplane RIR.

The multiplane models created for each RIR measurement case consisted of approximately 300 000 nodes, which equates to less than 2% of the number of nodes required for the 3D FDTD model. Hence, a very large computational saving is gained through use of the multiplane model. This is further reflected in comparative computation times. For the generation of a 1s RIR at audio rate, the multiplane models required a processing time of approximately 60s, therefore achieving a run-time reduction of $\sim$98% compared against the 3D model.

3. VIRTUAL RIR CALIBRATION AND HYBRIDISATION

3.1. 3D FDTD RIR Calibration

Once the 2D and 3D FDTD RIRs have been obtained they are processed such that they can be combined with the mid-high frequency geometric (GA-) RIRs generated in ODEON. In the case of 3D FDTD RIRs, this task was undertaken in accordance with the RIR matching procedure documented in [6]. The initial step involves calibrating each GA-RIR by equalising the total acoustic energy recorded at the corresponding reference receiver, positioned at a distance of 1m from each sound source in the geometric model, to a unity value. To this end, the total energy E_T of the GA reference impulse response captured at each reference receiver was calculated using,

$$E_T = \sum_{n=1}^{N} p^2[n] \tag{2}$$

where $p[n]$ is the pressure value recorded at temporal sampling instant n and N is the length of the impulse response in samples. To proceed, a constant K, which may be applied to reduce the total IR energy to unity, is calculated using the following relation:

$$K = \sqrt{\frac{1}{E_T}} \tag{3}$$

As such, three values of K were calculated and applied to the corresponding GA-RIRs through multiplication, thus appropriately calibrating the geometric model for each response case. The resulting GA-RIR signals were then high pass filtered to remove all spectral components below a cut-off frequency of 2kHz. This cut-off frequency was selected to ensure that the results obtained from FDTD modelling, applied to create the hybrid RIR, consisted of measurements well within the usable bandwidth of the numerical scheme computed.

In order to avoid reducing the effectiveness of the energy matching procedure between geometric and FDTD results, it was first necessary to process the 3D FDTD RIRs and remove D.C. components arising due to the nature of the excitation function and erroneous pressure recordings occurring at mid-high frequencies due to dispersion effects. As such, each 3D FDTD RIR was passed through a 2^{nd} order D.C. blocking filter and a low pass filter with a cut-off of 2kHz. Having done so, it was then possible to calibrate the 3D FDTD RIRs with the geometric results by multiplying each response with an energy matching parameter n, defined in [6] as:

$$n = (5.437 \times (Fs \times 10^{-3})) - 3.6347 \tag{4}$$

The spectrally complete hybrid RIRs were then created by summing the corresponding fully calibrated geometric and 3D FDTD impulse responses.

3.2. 2D Multiplane FDTD RIR Calibration

Due to dissimilarities of the laws governing energy decay in 2D and 3D FDTD schemes, as documented in [11], it was expected that the multiplane model would exhibit comparatively longer decay times. For this reason, calibration of reflection coefficients applied to the boundary of each 2D plane was necessary in order to obtain RIR reverberation times (RT_{60}) consistent with those generated in the 3D model. To this end, a series of simple cubic 2 x 2 x 2 m 3D FDTD lattices were constructed to obtain the $RT_{60,3D}$ of each cube with constant reflection coefficients, equal to those utilized in the 3D live room model, applied to all surfaces. The recorded $RT_{60,3D}$ values were then inserted to the Norris-Eyring for 2D RT_{60} rearranged to make the subject the boundary absorption value α:

$$\alpha = 1 - \exp\left(\frac{-\pi S ln(10^6)}{cLRT_{60,3D}}\right) \tag{5}$$

where S and L were set to the surface area and side length of the 2 x 2 m cross-section of the 3D cubic schemes. Using (5) it was possible to calculate appropriate boundary absorption values, and hence reflection coefficients, for the 2D multiplane model that corresponded to those applied in both the geometric and 3D FDTD models.

The RIRs simulated by the multiplane model were filtered in a similar manner to the 3D FDTD RIRs in order to remove D.C. components and frequencies above 2kHz. In addition, a further filtering stage was required to remove the effects of afterglow, as per the procedure documented in [12]. This was appropriate as the afterglow phenomenon in 2D numerical schemes acts to erroneously skew the magnitude of low frequency spectral components.

The final calibration stage involved matching the total energy present in the filtered multiplane RIRs to that of the corresponding calibrated 3D FDTD RIRs and, by extension, the geometric RIRs. This was carried out by applying (2) to the multiplane RIRs and multiplying each RIR by a matching constant K_n defined as follows:

$$K_n = \sqrt{\frac{E_{T,3Dn}}{E_{T,2Dn}}} \tag{6}$$

where $E_{T,2Dn}$ and $E_{T,3Dn}$ are the total energies of the 3D and 2D multiplane RIRs, respectively, for each measurement case $n = 1, 2, 3$. Having been calibrated by K_n the multiplane RIRs were then summed with geometric results to produce the complete 2D multiplane FDTD/geometric hybrid impulse responses. Currently, the correct calibration of the 2D multiplane RIRs relies on the generation of corresponding 3D RIRs for use as a reference. It is intended that this reliance will be removed in future work through development of an analytical means of matching the energy levels present in GA-RIRs and multiplane RIRs closely following the work presented in [6] and [13].

4. OBJECTIVE RESULTS

The results documented in this section are derived from the measured RIRs and those simulated by the 2D and 3D FDTD hybrid approaches. For the purposes of this study, analysis of RIRs is constrained to low frequency spectra in order to evaluate the results generated by the 2D multiplane and 3D modelling methods and then to compare both sets of results with real acoustic data.

4.1. Low Frequency Analysis

Three impulse responses, representative of the three measurement configurations applied in practice (see Table 1), were collected from each hybrid acoustic model. Figure 2 depicts a graphical comparison of the low frequency magnitude responses of the measured RIRs and the 2D multiplane and 3D FDTD hybrid RIRs in each measured case. For all cases, it is apparent that the magnitude of the low frequency response obtained in practice (denoted 'real') is consistently greater than that of either the virtual RIRs. This discrepancy is simply due to a difference in sound excitation strength applied in the real and virtual environments and was rendered negligible by means of a normalisation procedure prior to construction of material used during subjective testing later described. The magnitude responses generated by both hybrid models appear to show good agreement in overall energy levels which exposes the success of the energy matching procedure previously discussed.

The spectra presented for measurement cases 1 and 2 depict good agreement between low frequency components of measured RIRs and RIRs produced by both models below 150Hz. In this frequency range, comparable alignment of resonant peaks is observed, suggesting that strong axial modes are well represented by both acoustic models. Referring to case 1 in the range of 160-200Hz, it is apparent that the multiplane model does not accurately recreate the modal aspects present in both measured and 3D modelled spectra. This is potentially due to the inherent inability of the multiplane model to capture prominent oblique modes occurring in this frequency range. However, a similar disparity is not observed in case 2, where the multiplane model achieves a far better representation of the notch (190Hz) occurring in the measured response than the full 3D model. The cause of this result is yet to be investigated. Beyond 200Hz, in both cases 1 and 2, a reasonable correlation exists between the multiplane, 3D and measured RIR spectra in terms of spectral component positioning. Measured low frequency behaviour in case 3 is shown to be better represented by the synthesised 3D RIR than that of the multiplane models, however, agreement between all spectra is notable below 100Hz. In this case the 2D multiplane response exhibits a lack of clarity in terms of defined resonant components above 200Hz comparative to measured and 3D modelled results. In summary, the overall representation of low frequency characteristics possessed by synthesised 3D RIRs and measured RIRs attained by the multiplane model is very encouraging considering the achieved reduction in run-time compared to full 3D FDTD modelling (see section 2.3).

4.2. Global Reverberation and Early Decay Times

Reverberation time T_{30} and early decay times (EDT) were derived from the RIRs captured in the live room and the virtual models for each measurement scenario in accordance with ISO documentation [14]. As such, it was possible to calculate global values for both parameters by averaging the values returned for each measurement case applied in both models and in practice. Table 3 provides a review of the recorded parameter values in the 500Hz frequency octave. The additional 'Just Noticeable Difference' (JND) ranges were calculated by applying 5% JND [15] for EDT and 30% JND for T_{30} [6]. These percentage values refer to the maximum deviation from the true value beyond which the difference in decay times become perceptible. Hence, the JND measures provided offer insight to the subjective tolerance range for EDT and reverberation time with reference to the value calculated for each environment.

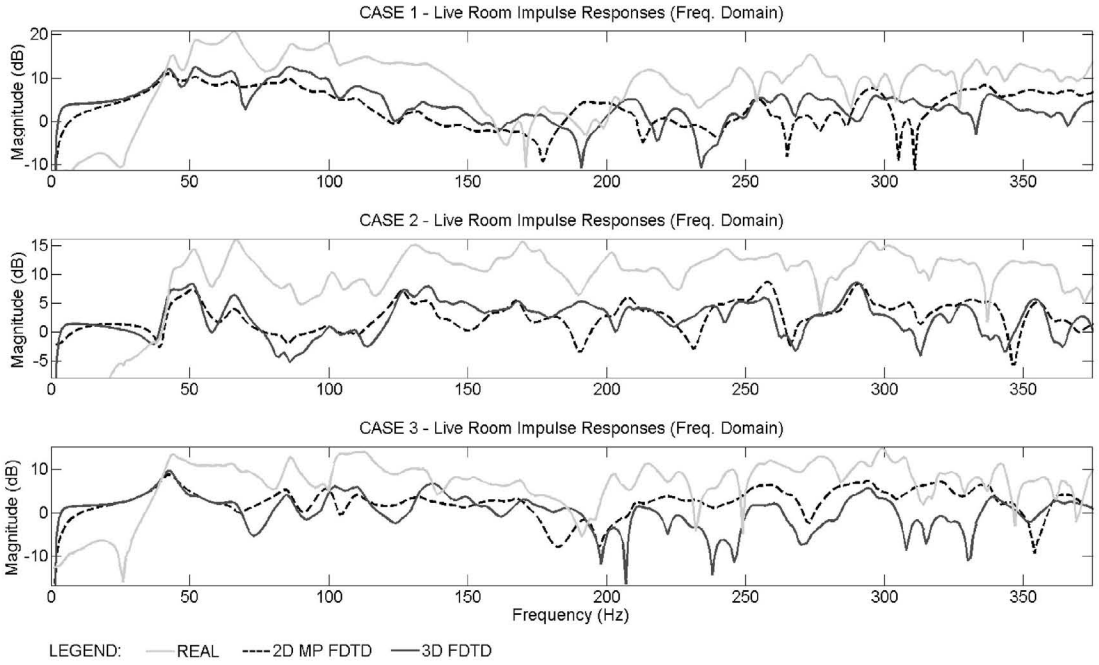


Figure 2: Comparison of magnitude spectra of real (solid grey), 2D Multiplane (black dashed) and 3D FDTD (solid black) RIRs at low frequencies for all three cases of source and receiver combinations.

	$T30_{500Hz}$ (s)	$T30$ JND Range (s)	EDT_{500Hz} (s)	EDT JND Range (s)
2D MP FDTD	0.38	0.27 - 0.49	0.26	0.25 - 0.27
3D FDTD	0.36	0.25 - 0.47	0.41	0.39 - 0.43
Real	0.44	0.31 - 0.57	0.42	0.40 - 0.44

Table 3: Global $T30_{500Hz}$ and EDT_{500Hz} values derived from RIRs measured in practice ('Real') and those rendered by the 2D multiplane and 3D FDTD hybrid acoustic models ('2D MP FDTD' and '3D FDTD' respectively).

As shown in Table 3, good agreement is exhibited between the 2D Multiplane and 3D FDTD model reverberation times with a minimal discrepancy of 0.2s. This demonstrates the effectiveness of the reflection coefficient matching procedure applied to the multiplane model (see section 3.2). A similar result is not observed for EDT where the difference between 2D and 3D models is much greater than JND tolerance values. This outcome is unexpected as energy levels, due to an impulse excitation, in 2D FDTD schemes should take longer to decay than those in 3D. Hence, it may be hypothesised that the values returned for the particular octave band under examination may not be representative of the overall EDT characteristics of the multiplane RIRs, however this claim requires further investigation. In addition, EDT is dependent on the distribution and amplitude of early reflections. In the case of the 2D multiplane model, low frequency reflection paths are represented only in planar cross-sectional areas as opposed to the volume of the modelled space in its entirety. As such, the temporal density of early reflections present in the multiplane FDTD RIRs is expected to be less than that of the 3D FDTD RIRs. This issue, which might also influence the EDT results presented in Table 3, remains to be examined in future work. Through comparison of the live room T_{30} with those of the models, it is apparent that reverberation times simulated in the virtual models lie within the JND range of the live room T_{30}, and hence the audible discrepencies should be negligible. With reference to the EDT measured in the live room, the 3D FDTD model is shown to produce the closest approximation of initial sound decay characteristics.

5. SUBJECTIVE TESTING

A simple preliminary listening test, described in the following, was constructed in order to support or disprove the following hypothesis:

"Auralisations generated by means of 3D FDTD/geometric and 2D Multiplane FDTD/geometric hybrid modelling will exhibit agreeable levels of similarity to auralisations rendered through use of measured RIRs in terms of perceived frequency response and reverberation."

5.1. Listening Test Material and Procedure

In review, a total of 9 RIRs were captured during the course of this study: 3 from measurements of the live room, and 3 from each of the hybrid models, representing 3 source and receiver configurations applied in practice. These RIRs were utilised to create aural-

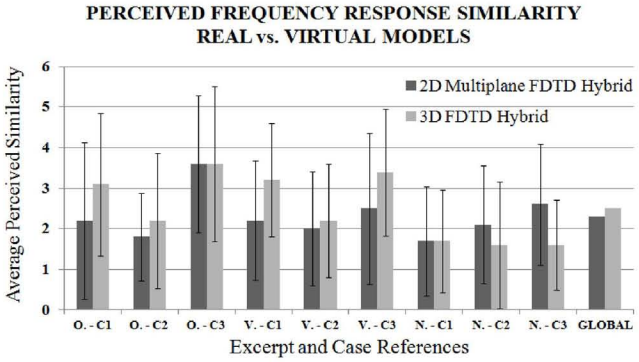


Figure 3: Average perceived similarity of frequency response between real and virtual auralisation for all excerpt and case combinations. Global averages are also shown.

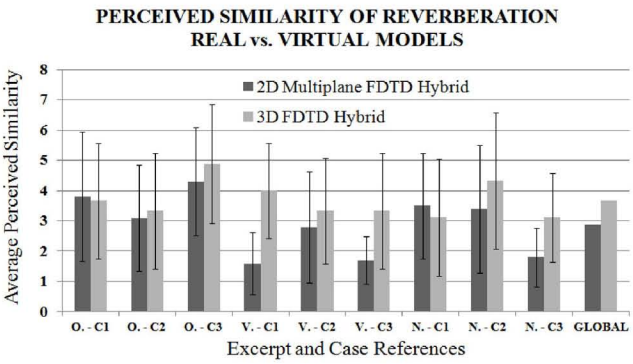


Figure 4: Average perceived similarity of reverberation between real and virtual auralisation for all excerpt and case combinations.

isations of the real and modelled spaces by convolving each with 3 anechoic audio excerpts: an orchestral recording (25s), an all-male vocal quartet (7s) and a pink noise burst (2s). This resulted in an audio test set of 27 auralisations, 9 of which were of the live room corresponding to the performance and capture of each anechoic excerpt as per all three cases of source and receiver placements. The remaining 18 auralisations were the virtual representations of the 9 auralisation scenarios as rendered by each hybrid acoustic model.

This auralisation set was arranged into 18 pairs such that each pair consisted of an auralisation rendered from live room measurements ('real') and the corresponding auralisation produced by means of one of the two hybrid acoustic models ('virtual'). To clarify, 9 test pairs compared the real auralisations with those rendered from 3D FDTD/geometric modelling and the remaining 9 compared real auralisations with those rendered from 2D Multiplane FDTD/geometric modelling.

During the test, subjects were presented with the auralisations in 'AB' pairs where, in each instance of the 18 pairs, 'A' was randomly selected as a real auralisation while 'B' was a virtual auralisation, or vice versa. Subjects were asked to listen to and compare the level of similarity between auralisation 'A' and 'B' with reference to frequency response and reverberation. Results were recorded by enabling subjects to rate the similarity of each perceptual parameter on a 7-point 'scale of similarity' on which a returned value of '0' corresponded to 'highly dissimilar' and '6' corresponded to 'highly similar'. This process was repeated for all 18 pairs of auralisations.

Due to the composition of the measured and rendered RIRs, the auralisations were rendered in Mono audio format and presented to subjects over headphones with both left and right channels producing the same signal. Volume levels remained consistent across all tests and the ordering of presented auralisation pairs was randomised for each subject.

5.2. Listening Test Results

The results provided by 10 test participants were combined and averaged to find the mean rating of similarity in terms of frequency response and reverberation for each model, case and excerpt combination. In doing so it was possible to compare the level to which each hybrid model compared with the auralisations of the live room and, hence, discern the effectiveness of the 2D multiplane FDTD hybrid model comparative to that of the 3D FDTD hybrid system, as per the test hypothesis.

Figure 3 displays the average perceived similarity of frequency response between both virtually modelled and real auralisations. For brevity, excerpts are denoted O., V., and N. for orchestra, voice and noise respectively. Likewise, cases 1-3 are denoted C1, C2 and C3. The error bars shown describe the standard deviation of recorded results from all tests with respect to the mean similarity values. As such, it may be observed that the perception, or rating, of similarity varied significantly between subjects for each auralisation comparison. However, from the mean values it is clear that, in terms of frequency response, the 3D and multiplane hybrid models performed comparably for O. - C3 and N. - C1. In contrast, the 3D hybrid model auralisations were perceived as most similar to the live room auralisations for the vocal excerpt in all measurement cases. This is most likely due to the content of the vocal excerpt which consisted of a 4-part male vocal ensemble and, therefore, was comprised of mainly low-mid frequencies. Referring back to Figure 2, it is shown that the multiplane responses at low frequencies best match those of the 3D model responses for measurement case 2. This is reflected in the results for the vocal auralisations suggesting that for auralisations, comprised of mostly low frequencies, perceptual similarities in frequency response may be more dependent on the placement and definition of resonant spectral components than initially thought. This interesting outcome is to be examined in future work in order to verify whether or not this is the case. Subjective comparison between virtual model auralisations and real auralisations rendered using the pink noise excerpt suggests that the multiplane model exhibits a higher level of realism than the 3D model (N. - C2 and C3) in terms of perceived frequency response. This outcome, which is converse to that demonstrated for the vocal auralisations, demonstrates that 2D multiplane modelling may be more appropriate than 3D FDTD modelling for particular auralisation purposes. Lastly, global test results noted in figure 3 appear to support the subjective test hypothesis. These global values were calculated by taking the mean value of all average similarities shown in figure 3 for each virtual model. In doing so, the influences of the audio excerpts and RIR measurement cases on the test results were bypassed to gain an overall measure of similarity. The discrepancy between these values for the 2D multiplane and 3D FDTD hybrid models is encouragingly small being in the region of 0.21.

In figure 4, the average perceived similarity between the live room auralisations and those produced using the two hybrid acoustic models is displayed in terms of reverberation. Considering first the comparison of real and virtual auralisations produced using the vocal excerpt, it may be observed that the 3D model auralisations consistently outperform those of the multiplane model with respect to perceived reverberance. As previously noted, this particular excerpt may expose the inaccuracies of the multiplane approach and, hence, the disparity between measured EDTs noted in Table 3 may be impacting these results. For the orchestral auralisations, discrepency between the reveberance of each model is shown to be small. The similarity ratings for each model are variable in the case of the pink noise excerpt auralisations, however global results (calculated analogously to those in figure 3) suggest a reasonably small comparative difference between modelling approaches in terms of perceived reverberance and, hence, provide partial support of the test hypothesis.

It is noted that the overall perceived similarity between model auralisations and auralisations rendered from practical RIR measurements is relatively low, as shown by global results for both frequency response and reverberation. This is expected considering the assumptions on which both hybrid modelling paradigms are based and the fact that significant acoustic phenomena, such as resonances in bounding surfaces, are not as yet possible to model. Moreover, particular aspects of the hybrid acoustic models could be refined in order to better represent a real acoustic environment. Such refinements would see the inclusion of approximated sound energy attenuation due to viscosity of air, frequency dependent boundary absorption characteristics at low frequencies and more accurate representations of sound source/receiver frequency response and directivity characteristics.

6. CONCLUSION

This paper provides an overview, and assesses the performance, of a recently devised 2D multiplane FDTD hybrid modelling paradigm applied to a realistic acoustic modelling scenario. Objective results obtained from RIR measurements demonstrate that this efficient multiplane approach possesses the potential to model realistic low frequency sound fields to a level of accuracy similar to that of 3D FDTD schemes while reducing run-times by approximately 98%. Results generated by conducting subjective listening tests support the claim that 2D multiplane and 3D FDTD hybrid models produce comparable levels of realism in rendered auralisations when compared against auralisations generated from practical measurements. With reference to the findings of this study, it is proposed that future work will aim to investigate and contrast the representation of early reflections in 2D and 3D FDTD acoustic modelling to better match RIRs resulting in each case. Additionally, the applicability of the mutliplane technique for acoustic simulation will continue to be assessed through a series of further subjective tests.

7. REFERENCES

[1] L. Savioja, "Real-time 3D finite-difference time-domain simulation of low- and mid-frequency room acoustics," in *Proc. of the 13th Int. Conf. on Digital Audio Effects (DAFx-10)*, Graz, Austria, 2010.

[2] N. Raghuvanshi, B. Lloyd, N. K. Govindaraju, and M. C. Lin, "Efficient numerical acoustic simulation on graphics processors using adaptive rectangular decomposition," in *Proc. of the EAA Symp. on Auralization*, Espoo, Finland, 2009.

[3] N. Raghuvanshi, R. Narain, and M. C. Lin, "Efficient and accurate sound propagation using adaptive rectangular decomposition," *IEEE Transactions on Visualization and Computer Graphics*, vol. 15, no. 5, pp. 789–801, 2009.

[4] A. Kelloniemi, V. Valimaki, and L. Savioja, "Simulation of room acoustics using 2-D digital waveguide meshes," in *Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Toulouse, France, 2006, pp. 163–168.

[5] L. Savioja and V. Valimaki, "Reducing the dispersion error in the digital waveguide mesh using interpolation and frequency-warping techniques," *IEEE Transactions on Speech and Audio Processing*, vol. 8, no. 2, pp. 184–194, 2000.

[6] A. Southern, S. Siltanen, D. T. Murphy, and L. Savioja, "Room Impulse Response Synthesis and Validation Using A Hybrid Acoustic Model," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 21, no. 9, pp. 1940–1952, 2013.

[7] S. Oxnard and D. Murphy, "Room impulse response synthesis based on a 2D multi-plane FDTD hybrid acoustic model," in *Proc. of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, New Paltz, NY, 2013.

[8] ODEON Website. (2014) Last Accessed 15th Feb. [Online]. Available: http://www.odeon.dk

[9] K. Kowalczyk and M. van Walstijn, "Room Acoustics Simulation Using 3-D Compact Explicit Schemes," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 19, no. 1, pp. 34–46, 2011.

[10] S. A. V. Duyne and J. O. Smith III, "Physical Modeling with the 2-D Digital Waveguide Mesh," in *Proc. of the Int. Conf. on Computer Music*, San Francisco, CA, 1993, pp. 40–47.

[11] J. Wells, D. Murphy, and M. Beeson, "Temporal matching of 2D and 3D wave-based acoustic modeling for efficient and realistic simulation of rooms," in *Audio Eng. Soc. (AES) Conv. 126*, Munich, Germany, 2009.

[12] J. Escolano, C. Spa, A. Garriga, and T. Mateous, "Removal of afterglow effects in 2-D discrete-time room acoustics," *J. Applied Acoustics*, vol. 74, no. 6, pp. 818–822, 2013.

[13] S. Siltanen, A. Southern, and L. Savioja, "Finite-difference time domain method source calibration for hybrid acoustics modeling," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Vancouver, Canada, 2013, pp. 166–170.

[14] European Standard Documentation, "Acoustics - Measurement of room acoustic parameters - Part 1: Performance spaces," *ISO 3382-1:2009*, 2009.

[15] N. Prodi and S. Velecka, "The evaluation of binaural playback systems for virtual sound fields," *J. Applied Acoustics*, vol. 64, no. 2, pp. 147–161, 2003.

EMBEDDING DISTANCE INFORMATION IN BINAURAL RENDERINGS OF FAR FIELD RECORDINGS

César Salvador, Shuichi Sakamoto, Jorge Treviño and Yôiti Suzuki

Advanced Acoustic Information Systems Laboratory,
Res. Inst. Electrical Communication and Grad. Sch. Information Sciences, Tohoku University
Sendai, Japan
{salvador@ais., saka@ais., jorge@ais., yoh@}riec.tohoku.ac.jp

ABSTRACT

Traditional representations of sound fields based on spherical harmonics expansions do not include the sound source distance information. As multipole expansions can accurately encode the distance of a sound source, they can be used for accurate sound field reproduction. The binaural reproduction of multipole encodings, though, requires head-related transfer functions (HRTFs) with distance information. However, the inclusion of distance information on available data sets of HRTFs, using acoustic propagators, requires demanding regularization techniques. We alternatively propose a method to embed distance information in the spherical harmonics encodings of compact microphone array recordings. We call this method the Distance Editing Binaural Ambisonics (DEBA). DEBA is applied to the synthesis of binaural signals of arbitrary distances using only far-field HRTFs. We evaluated DEBA by synthesizing HRTFs for nearby sources from various samplings of far-field ones. Comparisons with numerically calculated HRTFs yielded mean spectral distortion values below 6 dB, and mean normalized spherical correlation values above 0.97.

1. INTRODUCTION

The primary cues for distance perception are the intensity and the direct-to-reverberant energy ratio [1]. Recent studies suggest that listeners are also able to use binaural cues to determine the range of lateral sound sources for distances within 1 m [2, 3, 4, 5, 6]. Binaural cues can hence be used to determine directions and distances of nearby sound sources. However, it is difficult to include distance information on available far field HRTFs. The simplest approximation uses a head-sized sphere to model distance variations [7]. Better approximations require to solve an acoustic propagation problem using demanding regularization techniques [8, 9, 10].

We alternatively propose a method to edit distance information in the spherical harmonics encodings of distant sources. Our method is intended to make sounds appear closer or farther than their original distance during its binaural rendering (see Figure 1). At the recording stage, we assume sound fields captured by a compact spherical microphone array. At the reproduction stage, we rely on the use of a surrounding distribution of virtual secondary monopole sources rendered with far field HRTFs. A discrete distribution of this kind of virtual sources rendered with HRTFs can be understood as an array of virtual loudspeakers [11]. Hence, we refer to this reproduction scheme as the virtual loudspeaker approach. To match the sound field at the central area in the virtual loudspeaker positions to the field in the microphone positions, we perform spherical re-samplings based on spherical harmonics and

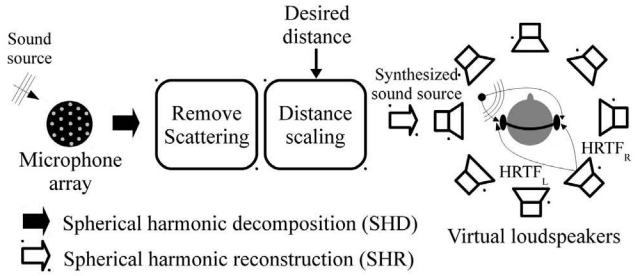


Figure 1: *Overview of the binaural synthesis method.*

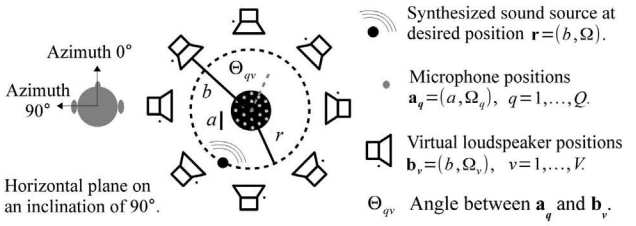


Figure 2: *Geometry used for the binaural synthesis method.*

distance manipulations based on Hankel functions. Binaural signals are finally rendered using the distance-edited encodings.

A top view of the assumed geometry is shown in Figure 2. A point in space $\mathbf{r} = (r, \theta, \phi) = (r, \Omega)$ is specified by its distance r, its inclination $\theta \in [0°, 180°]$, and its azimuth $\phi \in [0°, 360°]$. The listener's ears lie on an inclination $\theta = 90°$. The front direction lie on an azimuth $\phi = 0°$.

Section 2 overviews sound field analysis and binaural synthesis techniques. Section 3 overviews the synthesis of HRTFs for arbitrary positions from continuously available far field HRTFs. Section 4 describes the continuous formulation of our proposal. Section 5 evaluates our proposal in a practical scenario, where microphones and virtual loudspeakers are placed on spherical samplings. Conclusions are presented in Section 6.

2. BINAURAL AMBISONICS

2.1. Binaural rendering from spherical harmonics encodings

The Schmidt semi-normalized spherical harmonics, of order n and degree m, are denoted by $Y_{nm}(\theta, \phi) = Y_{nm}(\Omega)$. They form an orthonormal basis for the set of square-integrable functions on the

unit sphere $\mathbb{S}^2$. The sound pressure $f(\Omega)$ on the unit sphere is a function in this set. It can be expanded as [12]:

$$f(\Omega) = \sum_{n=0}^{\infty} \sum_{m=-n}^{n} f_{nm} Y_{nm}(\Omega), \tag{1}$$

where the coefficients f_{nm} define its spherical spectrum [12]:

$$f_{nm} = \int_{\Omega' \in \mathbb{S}^2} f(\Omega') Y^*_{nm}(\Omega') d\Omega'. \tag{2}$$

Eqs. (1) and (2) are respectively called the spherical harmonic reconstruction (SHR) and decomposition (SHD). A captured sound pressure field can thus be encoded with the SHD and decoded with the SHR. This defines the traditional High Order Ambisonics (HOA) format, a scalable way to render sound fields by decoupling the directions of the recording (Ω') and reproduction (Ω) setups.

Binaural reproduction of sound fields encoded by Eq. (2) is also possible. Encodings are decoded for a surrounding array of V virtual secondary sources using Eq. (1). The secondary source driving signals D_v derived in this way are then rendered with HRTFs H_v for the corresponding directions. Binaural signals B consist on superposing the resulting signals from all directions Ω_v:

$$B = \sum_{v=1}^{V} D_v H_v \alpha_v, \tag{3}$$

where D_v is decoded from existing encodings f_{nm} as follows:

$$D_v = \sum_{n=0}^{\infty} \sum_{m=-n}^{n} f_{nm} Y_{nm}(\Omega_v), \tag{4}$$

and the normalization factor α_v is applied to the virtual loudspeakers, so that they cover almost equal areas.

2.2. Distance manipulation of multipole encodings

The multipole expansion extends Eq. (1) to include distance information. The pressure $g(\mathbf{r}) = g(r, \theta, \phi)$ on a sphere of radius r can be expanded by [12]:

$$g(\mathbf{r}) = \sum_{n=0}^{\infty} \sum_{m=-n}^{n} g_{nm} h_n(kr) Y_{nm}(\Omega), \tag{5}$$

where h_n is the spherical Hankel function, and the coefficients g_{nm} can also be derived from the pressure $g(\mathbf{r}')$ on a sphere of different radius r', as follows [12]:

$$g_{nm} = \frac{1}{h_n(kr')} \int_{\Omega' \in \mathbb{S}^2} g(\mathbf{r}') Y^*_{nm}(\Omega') d\Omega'. \tag{6}$$

Eqs. (5) and (6) are the basis for the treatment of distances in sound field analysis. They have been applied to recording and reproduction technologies like Near Field Compensated High Order Ambisonics (NFC-HOA) [13]. These equations relates the pressure on a recording sphere of radius r' and the pressure on a reproduction sphere of radius r. Binaural rendering with virtual loudspeakers can also be done in a similar way to Section 2.1. However, spherical harmonics encodings cannot be easily converted into the NFC-HOA format, since this requires determining the reference distance r' established during recording. Furthermore, NFC-HOA systems seek accurate reproduction, while some recordings may be enhanced by making sounds appear closer or farther than their original distance.

3. USING HRTFS FOR CONTINUOUS DISTRIBUTIONS OF FAR SOURCES

In this section, the ideal case where HRTFs are continuously available for distant sound sources at a fixed radius is introduced. In this continuous case, the binaural synthesis of nearby sound sources is formulated as an acoustic propagation problem. We do not consider reverberant fields. Hence, we use the term *far field* to refer to spherical sound fields for which HRTFs hardly depend on distance, that is, to sound sources beyond 1 m distance from the listener's head [2, 3, 4, 5, 6].

The Helmholtz' principle of reciprocity allows to formulate the measurement of HRTFs as an acoustic radiation problem [14]. Two original sound sources are assumed to be located at the listener's ears, and a measurement sphere of radius b is centered on the listener. Here, all the sources of scattering from the head and torso of the listener, together with the original sound source, all of them constitute the source field. When torso is not considered, the head's radius r_h is defined as the smallest sphere's radius containing the head, hence containing the source field too.

Given an initial set of HRTFs denoted by $H(\mathbf{b}, k)$, measured on the sphere $\mathbf{b} = (b, \Omega)$ enclosing the head for a source field of wave number k, the HRTFs denoted by $\hat{H}(\mathbf{r}, k)$ on any other sphere $\mathbf{r} = (r > r_h, \Omega)$ containing the source field are completely defined by the simple source formulation [12]:

$$\hat{H}(\mathbf{r}, k) = \int_{\Omega \in \mathbb{S}^2} G(\mathbf{r}, \mathbf{b}, k) H(\mathbf{b}, k) d\Omega, \tag{7}$$

where $G(\mathbf{r}, \mathbf{b}, k)$ are the Green functions of wave number k characterizing the sound transmission in free space, from all monopole sources located at $\mathbf{b}$ to each desired position $\mathbf{r}$.

Multipole expansions of the Green function in Eq. (7) has been used to synthesize HRTFs for arbitrary positions from the initial set of HRTFs at a single radius [8, 9, 10]. Accurate synthesis is obtained following this approach. However, the source positions of the initial set need to be distributed almost uniformly on the sphere, for the radiation problem is formulated on the spherical harmonics domain. Otherwise, the multipole expansions requires regularization techniques according to particular geometries, for which an appropriate selection of the regularization parameter can become a demanding task.

4. DISTANCE EMBEDDING FOR HIGH ORDER AMBISONICS WITH BINAURAL RENDERING

In this section, our alternative proposal to embed distance information on recordings of distant sound sources is described.

4.1. Using far field recordings by a rigid continuous sphere

An alternative approach to the multipole expansion of the Green functions assumes a surrounding and continuous distribution of monopole secondary sources. The surrounding secondary sources are placed at the same radius $b > 1$ m as the initial set of far field HRTFs and binaurally rendered with them. This reproduction scheme is called the virtual secondary source approach [11]. The requirement of expanding the Green functions $G(\mathbf{r}, \mathbf{b}, k)$ is thus relaxed, and the signals $D(\mathbf{r}, \mathbf{b}, k)$ to drive the virtual secondary sources are computed instead. The driving signals are typically derived so as to match the sound pressure field due to sound sources in the far field, on a radius $r > 1$ m, using sound field

analysis techniques [15, 16]. However, sound field techniques typically decompose the sound field into plane waves, thus neglecting the distance-related effects, which may be important for a binaural rendering with high levels of realism. Our proposal follow the virtual secondary source approach considering the distance effects. In fact, we will derive the driving signals from far field recordings, but assuming point-like sources instead of plane waves. Therefore, in our proposal, the distance information can be further edited.

We next derive the signals to drive the continuous distribution of virtual secondary sources from a captured sound pressure field. By $D(\mathbf{r}, \mathbf{b}, k)$ we denote the driving signal of a virtual secondary source placed at $\mathbf{b}$, associated to a sound field generated by a sound source of wave number k placed at $\mathbf{r}$. In particular, we assume that the sound sources are on the same radius where the virtual secondary sources are continuously distributed ($r = b > 1$ m).

4.1.1. Spherical spectra of recordings and driving signals

On the recording side, we consider an ideal scenario where the pressure field is captured by a continuous, rigid and spherical sensing surface of radius a. In other words, the far field recordings, which we denote by $M(a, k)$, are available at an infinite number of points $\mathbf{a} = (a < b, \Omega')$. We characterize the recorded signals using the model of the acoustic scattering from the rigid sphere due to a point-like sound source. The total pressure on the surface of the rigid sphere reads [17]

$$S(\mathbf{a}, \mathbf{b}, k) = -\frac{1}{ka^2} \sum_{\nu=0}^{\infty} \frac{h_\nu(kb)}{h'_\nu(ka)} (2\nu + 1) P_\nu(\cos \Theta_{\mathbf{ab}}), \quad (8)$$

where $\Theta_{\mathbf{ab}}$ is the angle between the measurement point $\mathbf{a}$ and the source position $\mathbf{b}$, P_ν is the Legendre function, h_ν is the spherical Hankel function and h'_ν its derivative. In addition, we consider the recording spherical spectrum coefficients $S_{nm}(a, k)$, of order n and degree m, which reads [12]:

$$S_{nm}(a, k) = \int_{\Omega'} S(\mathbf{a}, \mathbf{b}, k) Y^*_{nm}(\Omega') d\Omega'. \quad (9)$$

On the virtual reproduction side, though, we assume driving signals whose spherical spectrum coefficients $D_{nm}(b, k)$ vanish for orders greater than N. Expansions of the driving signals in terms of spherical harmonics, evaluated in the secondary source directions Ω, are therefore defined by [12]

$$D(\mathbf{b}, k) = \sum_{n=0}^{N} \sum_{m=-n}^{n} D_{nm}(b, k) Y_{nm}(\Omega). \quad (10)$$

The spherical harmonics encodings are independent of the decomposing directions. Hence, the spherical spectra of the recording and virtual reproduction signals can be related by means of propagating filters from the radius a to the radius b.

4.1.2. Filters on the spherical spectrum

By F_n we denote the distance propagation filters. To derive F_n, we replace $D_{nm}(b, k)$ in Eq. (10) by the product of F_n with the spherical spectrum $S_{nm}(a, k)$ of Eq. (9). We then proceed to use the orthonormality property of spherical harmonics, and their addition theorem to decompose the Legendre polynomials into a sum of spherical harmonics products [12]. Assuming infinite recording points, it can be shown that the driving signals become

$$D(\mathbf{r}, \mathbf{b}, k) = \sum_{n=0}^{N} \frac{-F_n h_n(kb)}{ka^2 h'_n(ka)} (2n + 1) P_n(\cos \Theta_{\mathbf{rb}}), \quad (11)$$

where $\Theta_{\mathbf{rb}}$ is the angle between the source position $\mathbf{r}$ and the virtual secondary source position $\mathbf{b}$.

The filters F_n in Eq. (11) are chosen in such a way that they compensate for the distance effects. These filters therefore read

$$F_n(a, b, k) = -ka^2 \frac{h'_n(ka)}{h_n(kb)}, \quad (12)$$

whose factors compensate for the scattering effects introduced by the rigid sphere of radius a, and propagate the recordings on the radius a to the radius b where the secondary sources are. These filters are typically used to capture sound fields with rigid spherical microphone arrays [12, 18].

4.1.3. Distance-embedding filters

In addition, the theory of acoustic holography [12] allows to compute the near field compensation filters $\frac{h_n(kb)}{h_n(kr)}$ to estimate the pressure field at a new distance r. The driving signals for an arbitrary distance r can therefore be synthesized by applying the filters

$$F_n(a, r, k) = \frac{h_n(kb)}{h_n(kr)} F_n(a, b, k) = -ka^2 \frac{h'_n(ka)}{h_n(kr)} \quad (13)$$

to the spherical spectrum of the far field recordings.

The filters proposed in Eq. (13) do not depend anymore on the distance b of the original sound source, as long as the original source is placed beyond 1 m distance from the center of the listener's head. According to the acoustic radiation problem in Section 3.1, the minimum desired distance r that can be synthesized is the radius r_h of the smallest sphere containing the listener's head.

Application of Eq. (13) to sound fields recorded by compact microphone arrays and encoded with spherical harmonics enables the binaural rendering of sound sources at any distance $r > r_h$. We call this method the Distance-Editing Binaural Ambisonics (DEBA) hereafter.

5. APPLICATION OF DEBA

We proceed now to formulate and evaluate DEBA in a practical scenario, where microphones and virtual secondary sound sources are placed in almost regular samplings of the sphere.

5.1. Using spherical microphone arrays

In practice, a finite number Q of microphones is used on the recording side. The microphones are assumed to be placed at discrete points $\mathbf{a}_q = (a, \Omega_q)$ on the spherical surface. We denote each microphone signal by $M(\mathbf{a}_q, k)$, which arises from the discretization of $M(\mathbf{a}, k)$. We replace D_{nm} in Eq. (10) by the product of F_n with a quadrature over q of the recording spherical spectrum $\int_{\Omega'} M(\mathbf{a}, k) Y^*_{nm}(\Omega') d\Omega'$. We proceed to use the addition theorem of spherical harmonics [12] to deal with relative directions. The signals to drive the continuous distribution of secondary

sources, necessary to binaurally render nearby sound sources from the compact microphone array recordings, now read

$$D(\mathbf{r},\mathbf{b},k) = \sum_{n=0}^{N}(2n+1)F_n(a,r,k)\sum_{q=1}^{Q}P_n(\cos\Theta_q)M(\mathbf{a}_q,k)\beta_q, \quad (14)$$

where r is the desired distance, and Θ_q is the angle between the microphone at $\mathbf{a}_q$ and the virtual secondary source at $\mathbf{b}$. In particular, we considered almost constant integration quadratures β_q.

5.2. Using actual data sets of HRTFs

Measured sets of HRTFs are generally available for only some surrounding source positions at a fixed radius on the far field. Their spatial resolution is generally lower that the minimum audible angle of human auditory perception [19, 20]. To implement DEBA with such HRTF data set, an integral over the surface of the unit sphere similar to Eq. (7) need to be approximated by a weighted sum of a finite number of initial far field HRTFs. We refer to this kind of discrete distributions of secondary sources as virtual loudspeaker arrays. We therefore assume a finite number V of virtual loudspeakers placed at discrete points $\mathbf{b}_v = (b, \Omega_v)$ on the far field. We denote by $B(\mathbf{r},k)$ the binaural signals for a desired position $\mathbf{r}$. Hence, the binaural signals are synthesized as follows:

$$B(\mathbf{r},k) = \sum_{v=1}^{V} D(\mathbf{r},\mathbf{b}_v,k)H(\mathbf{b}_v,k)\alpha_v, \quad (15)$$

where α_v is the normalized quadrature weight that approximates the differential $d\Omega$ at each sampled point $\mathbf{b}_v$. In particular, we will use quadrature weights that are proportional to the area of each sampled point's neighborhood. We define the neighborhood of a sample as all points on the sphere that are closer to it than to other samples.

The driving signal $D(\mathbf{r},\mathbf{b}_v,k)$ in Eq. (15) arises from the discretization of Eq. (14). The driving signal for a virtual loudspeaker at $\mathbf{b}_v$ intended to render binaurally nearby sources from the microphone array recordings finally reads

$$D(\mathbf{r},\mathbf{b}_v,k) = \sum_{n=0}^{N}(2n+1)F_n(a,r,k)\sum_{q=1}^{Q}P_n(\cos\Theta_{qv})M(\mathbf{a}_q,k)\beta_q, \quad (16)$$

where $F_n(a,r,k)$ is the distance-embedding filter of Eq. (13), and Θ_{qv} now represents the angle between the microphone position $\mathbf{a}_q$ and the virtual loudspeaker position $\mathbf{b}_v$.

The filters F_n in Eq. (13) show high gains at low frequencies and high orders n, specially when using a rigid sphere of small radius a. In order to avoid low frequency distortion, spatial modes and frequencies are typically related. Hence, the reconstruction order N was chosen according to the wave number k and the scatterer size a as proposed in [20]:

$$N = \min(\lceil\frac{eka}{2}\rceil, \lfloor\sqrt{Q}-1\rfloor), \quad (17)$$

where e is the base of the natural logarithm and the number of microphones Q imposes the upper limit to the order.

Virtual loudspeakers should be placed on regular samplings of the sphere to avoid spatial aliasing. Regular spherical samplings, though, are only possible for the platonic solids. Among existing almost-regular samplings of the sphere, we have chosen the constructions based on the octahedron and the icosahedron. Icosahedral grids are constructed by subdividing the icosahedron's edges. They provide almost constant quadrature weights. In contrast, octahedral grids are constructed so to have octahedral rotation and inversion symmetry. They provide exact quadratures for numerical integration on the sphere [21] and, therefore, are suitable for computations with spherical harmonics. Figure 3 shows examples of icosahedral and octahedral grids, where dots indicate the positions of virtual loudspeakers and the lines enclose their neighborhoods.

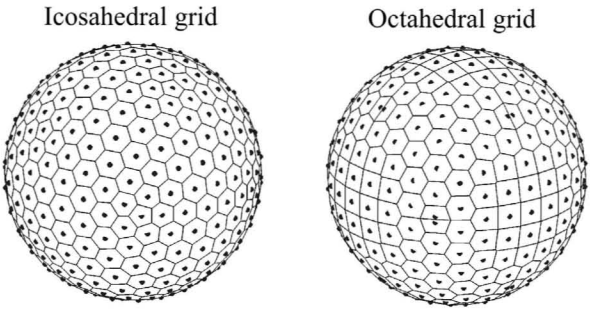


Figure 3: *Spherical grids to distribute the virtual loudspeakers.*

5.3. Conditions for the evaluation of the numerical accuracy

We need to know the effect of the number of virtual loudspeakers on the synthesis accuracy. For this purpose, microphone signals denoted by $M(\mathbf{a}_q,k)$ were characterized with Eq. (8) and the algorithm provided in [22]. The microphone signals correspond to 360 far field sound sources equiangularly distributed on the horizontal plane at a radius $b = 1.5$ m. Initial sets of far field HRTFs denoted by $H(\mathbf{b}_v,k)$ were computed numerically for a dummy head using the Boundary Element Method (BEM) [23]. The sound sources used to compute the far field HRTFs were arranged on icosahedral and octahedral grids, at a radius $b = 1.5$ m. Transfer functions for the whole binaural synthesis process, denoted by $B(\mathbf{r},k)$, were therefore characterized by using Eqs. (15) and (16), for several frequencies and desired positions in the horizontal plane. A reference set of near-field HRTFs, denoted by $H_{ref}(\mathbf{r},k)$, was also numerically computed using BEM. The resulting transfer functions for the whole binaural synthesis process were finally compared with the reference near-field HRTFs.

For each desired distance r, accuracy along azimuth θ was calculated by means of the spectral distortion (SD), defined by the logarithmic spectral distance between $H(\theta,f)$ and $B(\theta,f)$ [24]:

$$SD(\theta) = \left(\frac{1}{I}\sum_{i=1}^{I}\left(20\log_{10}\left|\frac{H_{ref}(\theta,f_i)}{B(\theta,f_i)}\right|\right)^2\right)^{\frac{1}{2}}. \quad (18)$$

Also for each desired distance r, accuracy along frequency f was calculated by the normalized spherical correlation (SC) between $H(\theta,f)$ and $B(\theta,f)$ [10]:

$$SC(f) = \frac{\sum_{j=1}^{J}H_{ref}(\theta_j,f)B(\theta_j,f)}{\left(\sum_{j=1}^{J}H_{ref}^2(\theta_j,f)\right)^{\frac{1}{2}}\left(\sum_{j=1}^{J}B^2(\theta_j,f)\right)^{\frac{1}{2}}}. \quad (19)$$

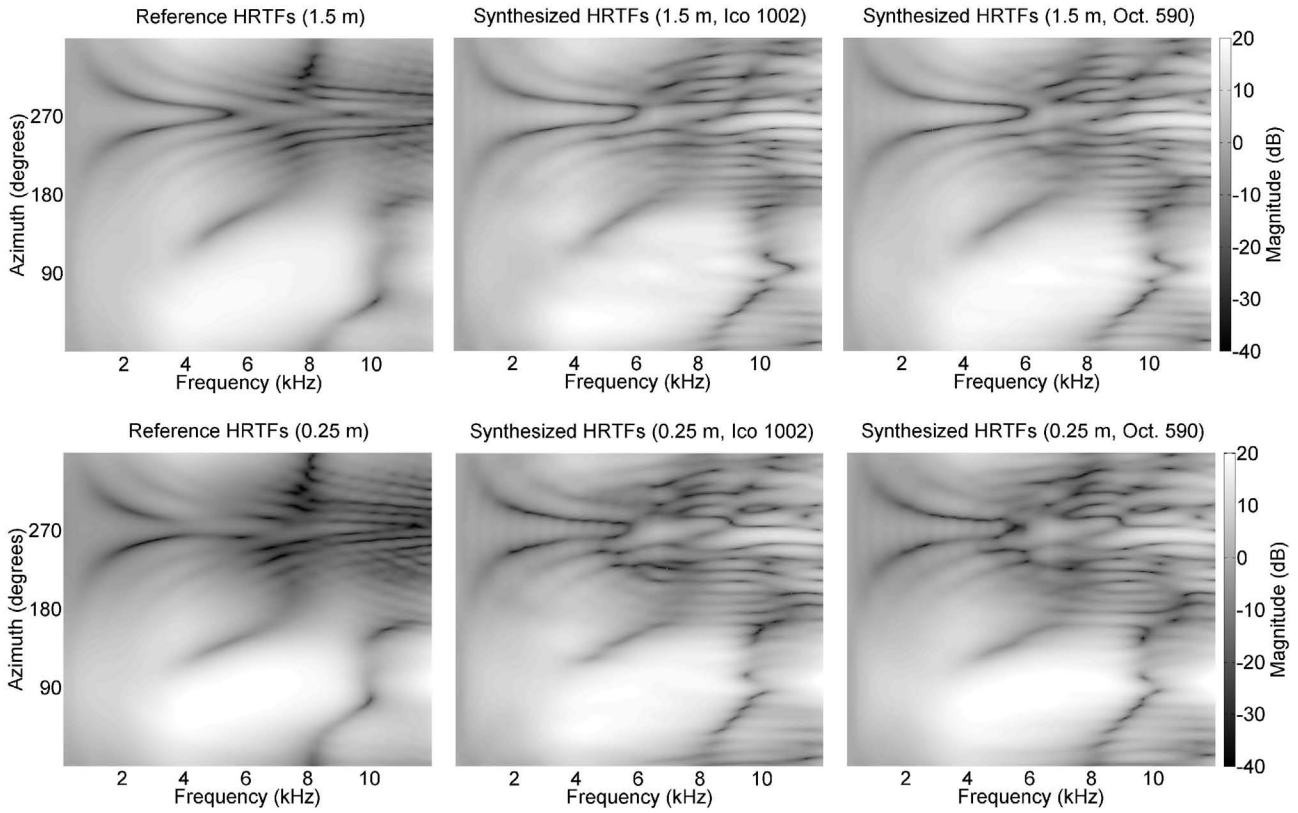


Figure 4: *Reference and synthesized HRTFs for distant (top) and nearby (bottom) sources on the horizontal plane (inclination of 90°). Reference HRTFs were numerically computed for a dummy head (left). We assumed 252 microphones and, hence, a spherical harmonics decomposition of order N = 14. Synthesis was performed with 1002 virtual loudspeakers on an icosahedral grid (middle) and 590 virtual loudspeakers on a octahedral grid (right), in both cases at a 1.5 m distance. These numbers of virtual loudspeakers correspond to the best accuracies (see Figure 5).*

We assumed microphones placed on a spherical scatterer of $a = 8.5$ cm radius, which we consider is the size of an average human head. According to [20], binaural synthesis in the entire audible frequency range, from 20 Hz to 20 kHz, would require an order $N = 43$, and therefore, a recording array of at least $Q = (43+1)^2 = 1936$ microphones. However, the practical number of microphones in existing compact arrays imposes a limited spatial bandwidth. At this stage, our evaluations were particularly focused on the recording setup available at the Research Institute of Electrical Communication in Tohoku University [25]. We therefore assumed $Q = 252$ microphones distributed in an icosahedral grid over the scatterer of $a = 8.5$ cm radius. This allowed for spherical harmonic expansions up to an order $N = 14$, and hence, accurate synthesis was only expected up to a spatial aliasing frequency of around 6.7 kHz.

5.4. Accuracy evaluation by computer simulations

Figure 4 shows some examples of HRTFs synthesized for the left ear and sound sources on the far (top panels) and near (bottom panels) regions. A visual comparison with the reference HRTFs on the right panels shows that the synthesis for sound sources placed on the same side of the ear (azimuth from 0°to 180°) can be performed with good accuracy up to around 8 kHz. Nevertheless, clearly decreasing accuracies appear for sound sources placed on the opposite side of the ear (azimuth from 180°to 360°). We noticed that the low-order spherical harmonics expansion does not yield a good approximation for the HRTFs. This was specially noticed for sound sources on the contralateral side of the ear, where signals of rapid variations along frequency and azimuths are caused by the head shadowing. In addition, discontinuity lines at some frequencies were due to the order limitation set by Eq. (17). Discontinuities are more prominent on the contralateral side and for desired distances near the head. On the other hand, slight decreasing accuracies appeared for distant and nearby sound sources of frequencies below 1.5 kHz. These particular observations suggested to focus the spectral distortion evaluations along azimuth on the contralateral side, and the spherical correlation evaluations along frequencies below 1.5 kHz.

Figure 5 shows the results of the numerical accuracy evaluation of the binaural synthesis performed with virtual loudspeakers on icosahedral (left panels) and octahedral (right panels) grids. The top panels show the mean values of the spectral distortion for sound sources on the opposite side of the left ear, along azimuths from 180°to 360°and frequencies below 8 kHz. The bottom panels show the mean values of the spherical correlation along all azimuths and frequencies below 1.5 kHz. Spectral distortions for contralateral sound sources yielded monotonically decreasing ac-

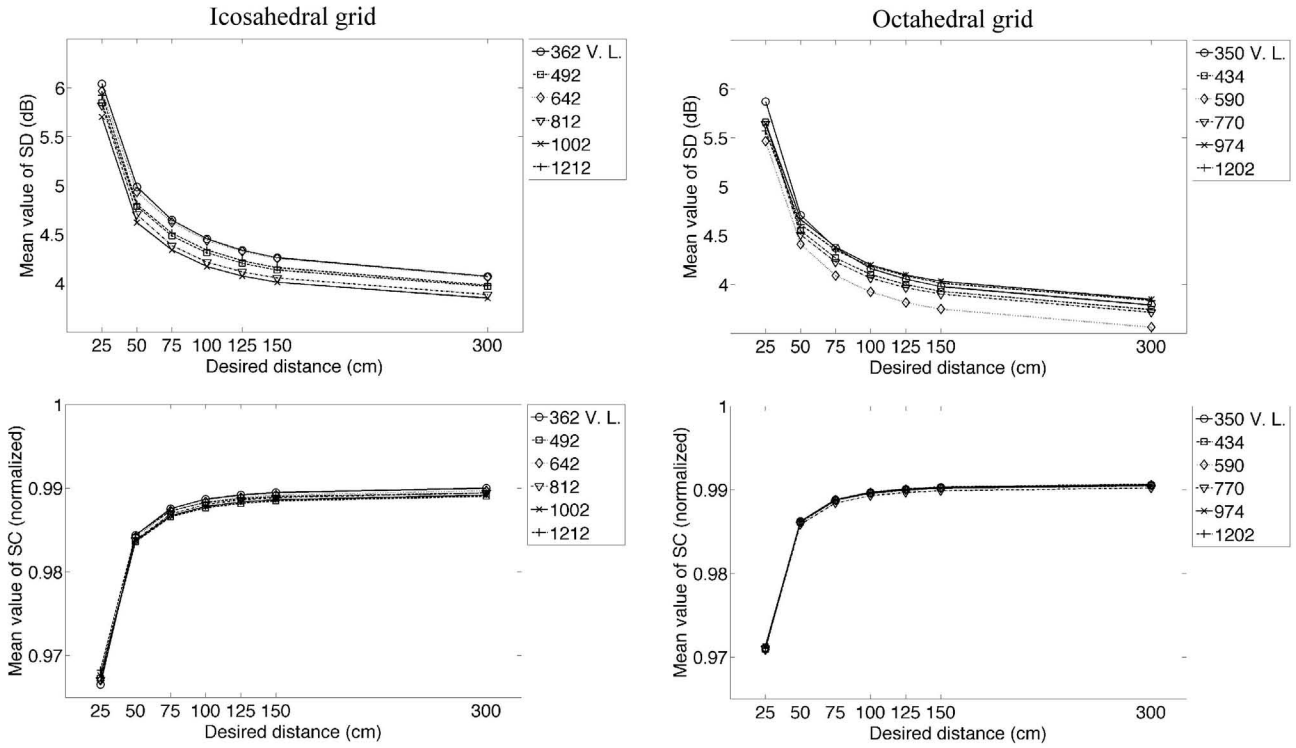


Figure 5: *Mean values of the spectral distortions on the contralateral side (top) and spherical correlations below 1.5 kHz (bottom) between the reference and synthesized HRTFs. Virtual loudspeakers (V. L.) were arranged on icosahedral (left) and octahedral (right) grids.*

curacies with decreasing desired distance. Regarding the number of virtual loudspeakers, the best accuracies were obtained using 1002 points in icosahedral grids and 590 points in octahedral grids, for which common minimums clearly appeared at all distances. On the other hand, spherical correlations for low frequency sound sources showed that accuracy is not affected by the number of virtual loudspeaker, but decreases monotonically with the desired distance. In general, evaluation using the spectral distortion yielded mean values below 6 dB, and using the spherical correlation, mean values above 0.97.

Our simulations were based on the addition theorem of spherical harmonics and, therefore, we did not consider the effects of matrix inversion based on regularization techniques, which are commonly applied in existing implementations of sound field encoding and decoding techniques [13]. In addition, typical sets of HRTFs are measured for non-uniform distributions of sound sources, making it necessary to use regularization techniques to match the virtual loudspeaker signals to sound field recordings. Although at this stage our evaluations were focused on the number of uniformly distributed loudspeakers, an extended study would require to add regularization techniques.

6. CONCLUSIONS

We proposed DEBA (Distance Editing Binaural Ambisonics), a method to synthesize the binaural signals at arbitrary sound source positions. We synthesized the binaural signals from the recordings made with microphones placed on the surface of a rigid sphere. For this purpose, we considered a surrounding array of virtual loudspeakers driven with head-related transfer functions. DEBA can accurately synthesize binaural signals due to sound sources placed on the horizontal plane. Accurate synthesis is possible up to the spatial aliasing limit imposed by the use of a finite number of microphones.

For evaluation, we relied on spherical harmonics encodings derived from the computer simulation of a compact, spherical microphone array. Transducers for both, recording and reproduction arrays were positioned in almost regular samplings of the sphere. Transfer functions for the whole process were characterized and compared with a set of near-field HRTFs computed numerically for a dummy head. Comparisons using the spectral distortion yielded mean values below 6 dB, and using the spherical correlation, mean values above 0.97. The accuracy cannot be improved by increasing the number of loudspeakers beyond the spatial aliasing limit imposed by the number of microphones. For lateral sources below 1 kHz, the accuracy decreased monotonically as the synthesized sound sources approaches the listener's head.

7. ACKNOWLEDGMENTS

This study was supported by Grant-in-Aid of JSPS for Scientific Research (no. 24240016), the Foresight Program for "Ultra-realistic acoustic interactive communication on next-generation Internet", and the Cooperative Research Project Program of RIEC Tohoku University (H24/A14). The authors wish to thank Makoto Otani for his efforts in developing the BEM solver used to generate the reference HRTF data.

8. REFERENCES

[1] Georg von Békésy, *Experiments in hearing*, McGraw-Hill, New York, NY, USA, 1960.

[2] M. Morimoto, Y Ando, and Z Maekawa, "On head-related transfer function in distance perception," in *Proceedings of the Congress of the Acoustical Society of Japan*, Japan, 1975, pp. 137–138, (in Japanese).

[3] Douglas S. Brungart and William M. Rabinowitz, "Auditory localization of nearby sources. hear-related transfer functions," *Journal of the Acoustical Society of America*, vol. 106, no. 3, pp. 1465–1479, Sept. 1999.

[4] David R. Moore and Andrew J. King, "Auditory perception: The near and far of sound localization," *Current Biology*, vol. 9, no. 10, pp. R361–R363, May 1999.

[5] Hae-Young Kim, Yôiti Suzuki, Shouichi Takane, and Toshio Sone, "Control of auditory distance perception based on the auditory parallax model," *Applied Acoustics*, vol. 62, no. 3, pp. 245–270, Mar. 2001.

[6] Pavel Zahorik, Douglas S. Brungart, and Adelbert W. Bronkhorst, "Auditory distance perception in humans: A summary of past and present research," *Acta Acustica united with Acustica*, vol. 91, pp. 409–420, 2005.

[7] Alan Kan, Craig Jin, and Andre van Schaik, "A psychophysical evaluation of near-field head-related transfer functions synthesized using a distance variation function," *Journal of the Acoustical Society of America*, vol. 125, no. 4, pp. 2233–2242, Apr. 2009.

[8] Ramani Duraiswami, Dmitry N. Zotkin, and Nail A. Gumerov, "Interpolation and range extrapolation of HRTFs," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, May 2004, vol. 4, pp. 45–48.

[9] Wen Zhang, Thushara Abhayapala, and Rodney A. Kennedy, "Insights into head-related transfer function: spatial dimensionality and continuous representation," *Journal of the Acoustical Society of America*, vol. 127, no. 4, pp. 2347–2357, Apr. 2010.

[10] Martin Pollow, Khoa-Van Nguyen, Olivier Warusfel, Thibaut Carpentier, Markus Müller-Trapet, Michael Vorländer, and Markus Noisternig, "Calculation of head-related transfer functions for arbitrary field points using spherical harmonics," *Acta Acustica united with Acustica*, vol. 98, no. 1, pp. 72–82, Jan. 2012.

[11] Jean-Mar Jot, Scott Wardle, and Veronique Larcher, "Approaches to binaural synthesis," in *Audio Engineering Society 105th Convention*, Paris, France, Sept. 1998, Audio Engineering Society.

[12] Earl G. Williams, *Fourier Acoustics: Sound Radiation and Nearfield Acoustical Holography*, Academic Press, London, UK, 1999.

[13] Jerome Daniel, "Spatial sound encoding including near field effect: Introducing distance coding filters and a viable, new ambisonic format," in *Audio Engineering Society Conference: 23rd International Conference: Signal Processing in Audio Recording and Reproduction*, Denmark, May 2003.

[14] Dmitry N. Zotkin, Ramani Duraiswami, Elena Grassi, and Nail A. Gumerov, "Fast head-related transfer function measurement via reciprocity," *Journal of the Acoustical Society of America*, vol. 120, no. 4, pp. 2202–2215, Oct. 2006.

[15] Markus Noisternig, Markus Sontacchi, Alois Musil, and Robert Holdrich, "A 3D ambisonic based binaural sound reproduction system," in *Audio Engineering Society 24th International Conference: Multichannel Audio, The New Reality*, Graz, Austria, June 2003, Audio Engineering Society.

[16] Sascha Spors and Jens Ahrens, "Generation of far-field head-related transfer functions using virtual sound field synthesis," in *German Annual Conference on Acoustics (DAGA)*, Mar. 2011.

[17] J. J. Bowman, T. B. A. Senior, and P.L.E. Uslenghi, *Electromagnetic and acoustic scattering by simple shapes*, Hemisphere, New York, NY, USA, 1987.

[18] Jens Meyer and Gary Elko, "A highly scalable spherical microphone array based on an orthonormal decomposition of the soundfield," in *Proceedings of the 2002 IEEE International Conference on Acoustics, Speech, and Signal Processing*, Orlando, FL, USA, May 2002, vol. II, pp. 1781–1784.

[19] A. W. Mills, "On the minimum audible angle," *Journal of the Acoustical Society of America*, vol. 30, no. 4, pp. 237–246, Apr. 1958.

[20] Wen Zhang, Mengqiu Zhang, Rodney A. Kennedy, and Thushara Abhayapala, "On high-resolution head-related transfer functions measurements: an efficient sampling scheme," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 20, no. 2, pp. 575–584, Feb. 2012.

[21] V. I. Lebedev, "Quadratures on a sphere," *{USSR} Computational Mathematics and Mathematical Physics*, vol. 16, no. 2, pp. 10 – 24, 1976.

[22] Richard O. Duda and William L. Martens, "Range dependence of the response of a spherical head model," *Journal of the Acoustical Society of America*, vol. 104, no. 5, pp. 3048–3058, Nov. 1998.

[23] Makoto Otani and Shiro Ise, "Fast calculation system specialized for head-related transfer function based on boundary element method," *Journal of the Acoustical Society of America*, vol. 119, no. 5, pp. 2589–2598, May 2006.

[24] Takanori Nishino, Naoya Inoue, Kazuya Takeda, and Fumitada Itakura, "Estimation of HRTFs on the horizontal plane using physical features," *Applied Acoustics*, vol. 68, pp. 897–908, Feb. 2007.

[25] Shuichi Sakamoto, Satoshi Hongo, R. Kadoi, and Yôiti Suzuki, "SENZI and ASURA: new high-precision sound-space sensing systems based on symmetrically arranged numerous microphones," in *Proceedings of the Second International Symposium on Universal Communication*, 2008, pp. 429–434.

Proc. of the EAA Joint Symposium on Auralization and Ambisonics, Berlin, Germany, 3-5 April 2014

COMPARISON OF SPATIAL AUDIO TECHNIQUES FOR USE IN STAGE ACOUSTIC LABORATORY EXPERIMENTS

Iain Laird[*]
Digital Design Studio
Glasgow School of Art, UK
I.Laird1@gsa.ac.uk

Damian Murphy
Audio Lab Department of Electronics
University of York, UK
damian.murphy@york.ac.uk

Paul Chapman
Digital Design Studio
Glasgow School of Art, UK
p.chapman@gsa.ac.uk

ABSTRACT

Real-time auralisation systems are increasingly being used by researchers aiming to observe how particular stage and auditorium configurations affect a musician's performance technique. These experiments typically take place in controlled laboratory conditions equipped with auralisation systems capable of reproducing the 3D acoustic conditions of a performance space in response to a performing musician in real-time. This paper compares the performance of First Order Ambisonics and Spatial Impulse Response Rendering in terms of both objective measurements and subjective listening tests. It was found that both techniques spatialised single reflections with similar accuracy when measured at the sweet spot. Informal listening tests found that the techniques produced very similar perceived results both for synthesised impulse responses and for measured stage acoustic impulse responses.

1. INTRODUCTION

In order to investigate specific subjective effects of stage acoustic conditions on a performing musician, it is necessary to introduce musician test subjects into known stage environments and allow them to play in the space, noting their subjective reaction to specific objective variables. The most straightforward way of providing this environment is to perform the experiments in existing performance spaces. However, gaining access to performance spaces can be costly and often it is not possible to control specific aspects of the acoustic response. Therefore, the alternative is to develop a laboratory system which is capable of presenting a test subject with controlled acoustic conditions allowing, for example, a musician to play in a virtual version of a performance space.

Such research has been emerging since the 1980's with Gade [1] pioneering an approach using electronic delays and reverberation chambers to create virtual versions of concert halls and investigating specific phenomena experienced by musicians. Digital audio and spatial audio techniques have moved on significantly since then and it is now possible to provide a listener with a much more accurate 3D simulation of a soundfield which can be adjusted to allow certain phenomena to be investigated in more detail.

[*] This work was supported by Arup Acoustics. Thanks to all the listening test volunteers and the Glasgow City Halls for allowing access to the venue

Recently, a number of authors have made use of real-time convolution [2] and Ambisonic recording/playback techniques [3] in order to emulate stage acoustic conditions for a performing musician in real-time. The use of First Order Ambisonics (FOA) provides a convenient and efficient way of capturing, analysing, transforming and recreating 3D soundfields in laboratory conditions. However, a real-time auralisation system capable of reproducing an accurate and natural sounding acoustic environment continues to present a significant challenge. FOA-based systems, are known to have a limited spatial resolution and due to the highly correlated nature of the loudspeaker signals, can often result in blurred and coloured reproduction of sound sources at the sweetspot [4]. Other techniques, such as Higher order Ambisonics (HOA) [5] have recently been used to increase the spatial accuracy of an auralised soundfield, however the lack of widespread availability of HOA microphones can in some cases prohibit the use of measured impulse responses.

Spatial Impulse Response Rendering (SIRR) [4], is a more recently developed spatial audio technique which is capable of providing detailed directional analysis, complex modification and reproduction of 3D impulse responses over arbitrary loudspeaker arrays. It is a perceptually motivated approach which analyses a 3D impulse response for physical properties that will transform into human auditory localisation cues. It then synthesises appropriate loudspeaker feeds aiming to recreate a natural sounding soundfield with the equivalent spatial impression. SIRR provides an attractive alternative to FOA-based real-time auralisation systems as it is possible to manipulate and analyse elements of an impulse response in much finer detail.

This paper presents a comparison of SIRR and FOA-based real-time auralisation systems for use in the context of stage acoustic laboratory tests. A series of objective and subjective tests were performed to indicate which technique is more appropriate for these types of experiments. The tests aimed to objectively compare the spatialisation quality of auralised impulse responses using each technique by measuring at the sweetspot of an auralisation system. Subjective tests were also carried out to determine if there were any audible differences using each technique when auralising various target acoustic conditions. If a SIRR-based system was demonstrated to produce similar or better objective and subjective results than an FOA-based system then it would be an

initial indication that using SIRR is a viable option for stage acoustic experiments and therefore the various advantages it has can be exploited in the future.

The structure of the paper is as follows: The first section begins by briefly describing the typical architecture of a real-time auralisation system used in stage acoustic laboratory tests. It will then describe the operation of SIRR for both analysis and re-synthesis of 3D impulse responses and compare the operating performance against FOA based systems using a series of objective tests. Finally, the paper will report on an informal listening test which aimed to ascertain if naive listeners could detect any subjective differences between 3D soundfields recreated using SIRR and FOA.

2. REAL-TIME AURALISATION SYSTEMS

A real-time auralisation system measures the direct sound created by a musician which is then processed by a computer capable of performing real-time convolution of the direct sound with a 3D impulse response of a performance space. In FOA-based systems, the direct sound is convolved with the four channels of a B-format impulse response representing the target space. The processed audio is passed to a decoder matrix which produces a number of speaker feeds which play back the auralised sound of the performance space back to the musician over a loudspeaker array. Figure 1 shows the layout of a typical FOA-based real-time auralisation system. A SIRR-based system operates in a similar way with the exception that the impulse response has been decomposed into a number of impulse responses (one per loudspeaker) and so there is no need for a decoder stage.

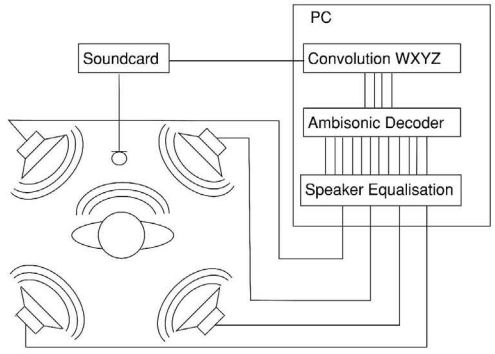


Figure 1: *System diagram of a typical FOA-based real-time auralisation system (only 4 loudspeakers are shown for clarity)*

A real-time auralisation system similar to that shown in figure 1 was constructed in the Arup-DDS SoundLab situated in Glasgow, UK. The SoundLab is an acoustically controlled space with dimensions 4m (l) x 6m (w) x 2.5m (h). The average T20 of the space at 500Hz is approximately 0.15s. The SoundLab has a measured background sound pressure level (L_{AF90}) of approximately 18dBA. The SoundLab features a 16-channel 3D loudspeaker array comprising of three rings of Yamaha MSP5A active loudspeakers arranged in a 4 - 8 - 4 arrangement as shown in figure 4. The speaker system was equalised to ensure a flat frequency response and an equal level contribution at the sweetspot.

For the purposes of stage acoustic auralisation it is neither necessary or practical to auralise the direct sound or the floor reflection as they are produced by the instrument and solid floor respectively. By silencing any audio before the first required reflection and then truncating by a value equal to the system latency, the effective latency of the system can be minimised. Typically, the direct sound is measured using directional microphones positioned close to the musician in order to minimise unwanted feedback. In this particular system the direct sound is measured using a single directional microphone, however further improvements could be made by using multiple microphones positioned around the musician to ensure the complex, time-varying radiation characteristics of the instrument are captured.

3. SPATIAL IMPULSE RESPONSE RENDERING

Spatial Impulse Response Rendering (SIRR) is a spatial audio technique which allows a 3D soundfield to be rendered to an arbitrary speaker layout [4]. The technique involves the analysis of a soundfield to obtain its directional properties and subsequent synthesis of the resulting diffuse and non-diffuse cues to recreate a perceptually equivalent soundfield. A 3D soundfield can be measured with an Ambisonic microphone and analysed in the time-frequency domain to produce a pressure signal with accompanying meta-data carrying information regarding the direction of arrival and diffuseness of each time-frequency element. The meta-data is then used to reconstruct the soundfield using amplitude panning techniques. SIRR has found many different applications including soundfield analysis [6], high quality auralisation of room acoustics [4] and parametric spatial audio effects [7].

3.1. Analysis

The direction of arrival of a sound can be estimated by obtaining the active sound intensity which is a product of the sound pressure $p(t)$ and particle velocity vector $u(t)$. This describes the transfer of energy of the soundfield and therefore the opposing vector will describe the direction of arrival of the sound. This can be achieved by analysing the B-format output of an Ambisonic microphone where the omnidirectional signal, $W(t)$, is assumed to be proportional to the pressure $p(t)$. The remaining orthogonal, figure-of-eight pressure-gradient signals $X(t)$, $Y(t)$ and $Z(t)$ can be considered as proportional to the components of the particle velocity $u(t)$. Therefore the active intensity can be obtained using equation (1) and the direction of arrival found using equation (2) and (3) giving azimuth and elevation respectively. Where $\acute{X}(\omega) = (X(\omega)e_x + Y(\omega)e_y + Z(\omega)e_z)$, "$*$" denotes complex conjugation and $Z_0 = \rho_0 c$ is the acoustic impedance of air.

$$I_\alpha(\omega) = \frac{\sqrt{2}}{Z_0}\Re\{W^*(\omega)\acute{X}(\omega)\} \tag{1}$$

$$\theta(\omega) = \tan^{-1}\left[\frac{-I_y(\omega)}{-I_x(\omega)}\right] \tag{2}$$

$$\phi(\omega) = \tan^{-1}\left[\frac{-I_z(\omega)}{\sqrt{(I_x^2(\omega) + I_y^2(\omega))}}\right] \tag{3}$$

Diffuseness can be estimated by obtaining the proportion of sound energy contributing to the net transport of energy and can be calculated using equation (4). This produces a value between 1 and 0 for each time-frequency element characterising the sound as either diffuse or non-diffuse. By multiplying the audio signals

by $\sqrt{\psi}$ or by $\sqrt{1-\psi}$ the audio signals can be separated into diffuse and non-diffuse signals respectively and re-synthesised with appropriate spatial audio techniques as described below.

$$\psi(\omega) = 1 - \frac{2Z_0\|\Re\{W^*(\omega)\acute{X}(\omega)\|}{|W(\omega)|^2 + |\acute{X}(\omega)|^2/2} \tag{4}$$

3.2. Synthesis

Synthesis of the audio signals takes place in the frequency domain by transforming the audio signals using the Short Time Fourier Transform (STFT) and applying the meta-data obtained in the analysis to the audio signals before using the Inverse Short Time Fourier Transform (ISTFT) to produce the output time-domain audio signals. The analysis-resynthesis process is shown in Figure 2.

Non-diffuse sound sound synthesis aims to reproduce coherent reflections as point-like sources and is typically implemented using Vector Base Amplitude Panning (VBAP) [8]. VBAP is an amplitude panning technique that allows sounds to be panned around a periphonic loudspeaker array using vector calculation to determine the level of a local triplet of loudspeakers. The diffuse sound synthesis aims to recreate the reduced interaural coherence produced by the diffuse sound energy. This is achieved by decorrelating the sounds identified as being diffuse and distributing equally to each loudspeaker. Decorrelation can can be implemented in a number of different ways however in this study decorrelation via time-varying phase randomisation was used which has been reported to give acceptable results when re-synthesising impulse responses [9].

High quality implementations of SIRR makes use of all B-format signals, applying the meta-data to a set of decoded signals for each loudspeaker using virtual microphone principles. This has been observed to provide better directional separation, natural decorrelation and overall higher audio quality. The directivity factor has been found to produce favourable results when set as a dipole microphone pattern and angled towards each loudspeaker [10]. It has also been shown that the use of virtual microphones can affect the correct reproduction of energy in both diffuse and non-diffuse sound. It has been demonstrated that it is possible to apply correction gains to ensure the correct ratio of diffuse and non-diffuse components [10, 9].

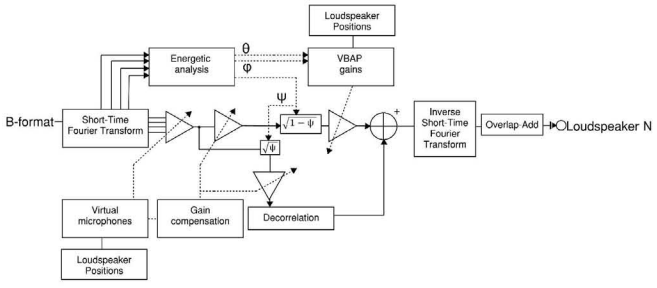


Figure 2: *System diagram of synthesis technique shown for a single loudspeaker channel*

The synthesis phase of SIRR produces a time-domain impulse response signal per loudspeaker channel which can then be utilised by a multichannel convolving reverberator to recreate the acoustic conditions of the target space at the sweetspot.

4. OBJECTIVE ANALYSIS

The intensity vectors obtained in the analysis phase (as described in section 3.1) can also be used to evaluate characteristics of 3D impulse responses. When a coherent reflection or plane wave is analysed using SIRR, the intensity vectors all tend to point in the direction of arrival at the time of arrival. Conversely, in diffuse (or reverberant) conditions, the intensity vectors are distributed in a more stochastic fashion. Conditions measured on stage typically consist of a number of coherent early reflections followed by a diffuse reverberant decay. Therefore, observing the angular distribution of intensity vectors in an impulse response will allow the presence of reflections to be detected [6]. Furthermore, the mean direction of the intensity vectors will indicate the direction of arrival of each reflection.

Spherical variance, σ (5), can be used to assess the angular distribution of the intensity vectors for each time step. It is defined as the magnitude of the mean resultant vector, $S(t)$ (6) where $\|\cdot\|$ determines the magnitude of the enclosed vector. The intensity vectors $I_i(t)$ are normalised prior to calculation of $S(t)$ to provide more robust results as demonstrated in[6].

$$\sigma(t) = 1 - \|S(t)\| \tag{5}$$

$$S(t) = \frac{1}{X}\sum_{i}^{X} I_i(t) \tag{6}$$

Where $I_i(t)$ is the i^{th} frequency band for time step t and X is the total number of frequency bins used in the analysis [6, 11]. The mean angular direction for each time frame can be subsequently computed using equations (2) and (3). For stage acoustic impulse responses, σ will be close to zero when $t = 0$ (as the direct sound arrives) and will increase quickly to a maximum value less than one as the impulse response becomes increasingly diffuse with time. The arrival of a coherent reflection will produce a localised trough in this response with a magnitude dependent on the nature of the reflection.

When using a short time window and low hop size in the time-frequency analysis, a single reflection may be identified a number of times with slightly different results per iteration resulting in a number of points representing the direction of arrival per reflection. A Gaussian Mixture Model (GMM) can be used to identify clusters of these points when arranged by diffuseness and direction of arrival. The component-mean (centre) of the cluster can then be used to estimate a direction of arrival.

5. TEST METHODOLOGY

In order to compare the performance of each spatialisation technique in the context of stage acoustic laboratory experiments, a real-time auralisation was set up in the SoundLab using known impulse responses rendered using FOA or SIRR. The impulse response of the auralised space within the SoundLab was measured at the sweetspot in order to emulate a musician using the space. This was achieved by positioning a directional loudspeaker and ambisonic microphone in the sweet spot of the loudspeaker array to represent the musician's instrument and head respectively.

The impulse responses used in the test consisted of either a single synthesised reflection arriving from a single direction or a measured stage impulse response obtained during a survey of the Grand Hall, Glasgow City Halls. The stage impulse response

was obtained by measuring in the venue using a Genelec 1029A Active loudspeaker and Soundfield ST350 Ambisonic microphone arranged in a manner emulating the instrument and head of a musician respectively. The apparatus was positioned down-stage right approximately 4m away from a nearby side wall. The average mid-frequency reverberation time (T30, 500Hz) at this position was found to be approximately 1.75 seconds. The synthesised impulse response consisted of a single, non-diffuse reflection with a time delay of 60ms relative to the direct sound which was panned using ambisonic panning techniques to various angles of azimuth ($0°$, $\pm$ $60° \pm 90°$, $\pm$ $110°$). The amplitude of this reflection was set to be -6dB below the direct sound.

Both the measured and synthesised FOA impulse responses were processed as described in section 3 to obtain an impulse response for each loudspeaker in the SoundLab. The analysis and re-synthesis was implemented in the time-frequency domain using a MATLAB script. An STFT was used with a 16-sample Hanning window (with 16 samples of zero padding) and a hopsize of 4 samples. Signal reconstruction utilised an ISTFT alongside the OverLap-Add method (OLA) in order to obtain as close to perfect reconstruction as possible. The window size was chosen to ensure that the highly transient nature of the impulse response signals were preserved. It is noted however that the small window size will require a compromise in terms of frequency resolution due to the inherent trade-off between time and frequency resolution of the STFT.

A Genelec 1029A Active loudspeaker was mounted on a tripod at a height of 100cm above the floor (height to top of main driver) and a radial distance of approximately 50cm from the receiver to the centre of the loudspeaker. The receiver was an Soundfield ST350 Ambisonic microphone which was located in the sweet-spot at a height of 130cm shown in Figure 3. Impulse responses were measured using a 10-second, logarithmically swept sinusoidal signal generated in MATLAB. The sine sweep was played through the loudspeaker and measured simultaneously at a sampling frequency of 44.1kHz and a 32-bit floating-point bit depth. Impulse responses were extracted from the recorded sine sweeps by convolution with the inverse of the original sweep as demonstrated by Farina [12].

The direct sound from the loudspeaker was measured using an M-Audio Luna cardioid condenser microphone positioned on axis between low and high frequency drivers of the loudspeaker at a 10cm distance. As with previous experiments [2], the direct sound in this experiment was measured using a single directional microphone. This is a noted simplification as a musical instrument exhibits complex and time-varying radiation characteristics which cannot be measured using a single microphone. The sound source used to measure the stage impulse responses is identical to that used in this experiment therefore the error in using a single microphone to measure the direct sound is minimised. It will be necessary for future stage acoustic laboratory tests to consider the impact of this.

The direct sound was input into the auralisation system as described in section 2. When auralising using FOA, real-time convolution was performed in Max MSP using two, 4 SIR2 VST convolution engines, the resultant convolved audio was decoded to the 16-channel loudspeaker array in the SoundLab using a Gerzonic Decopro ambisonic decoder set to a "Max-RE" type for all frequencies. When auralising using SIRR, the number of convolution objects increased to 16 (one per loudspeaker) and the ambisonic decoder is removed.

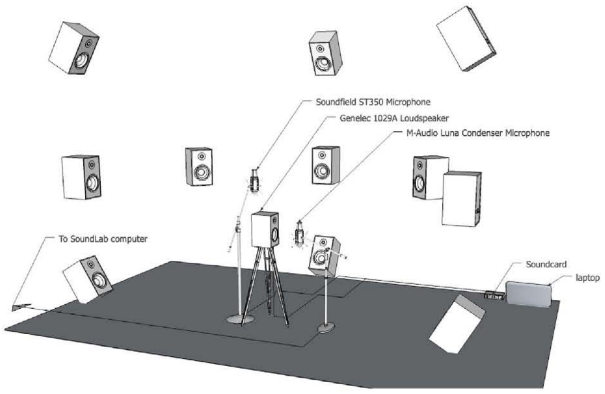


Figure 3: *Model of experimental set up in SoundLab (loudspeakers on one side have been omitted for clarity)*

6. OBJECTIVE RESULTS

Figure 4 shows an example set of objective results for an auralised reflection arriving at an angle of $-90°$ (anticlockwise), 60ms after the direct sound at a level of -6dB below the direct sound. The omnidirectional amplitude response of this scenario is shown for both FOA and SIRR. The direct sound occurs shortly after $t = 0$ and the reflection clearly arrives at $t = 60ms$. The decay seen after each of these events is caused by the SoundLab acoustic response. Overlaid on this plot is the spherical variance for both techniques. It can be seen the spherical variance is very similar for the first 50ms which starts at a value of zero when $t = 0$ and increases rapidly during the SoundLab acoustic decay. After $t = 50ms$, there are clear differences for each technique when the reflection arrives. It can be seen there is a sharp reduction in spherical variance when the reflection arrives which is of greater amplitude when using SIRR than FOA.

Figures 5(a) and 5(b) show the angle of the mean resultant vector for the first 0.2 seconds of the synthesised impulse response using FOA and SIRR respectively. For clarity, the plots show the analysis for parts of the impulse response that are below a mean diffuseness value of 0.55. The smaller dots show the analysed angle of arrival while the larger dots show the GMM estimate. The expected angle of arrival is shown by a cross which in this case is positioned at $-90°$.

When using the GMM estimation method to estimate the angle of arrival from clusters of points, errors can be introduced into the estimation as the location of a component-mean is influenced by nearby clusters. These clusters can be created by periods of silence or other nearby reflections. In this case, the adjacent clusters are caused by the room acoustic response of the SoundLab. Therefore for each GMM estimation, three component-means were calculated. This is to ensure that the auralised reflection is estimated by one component-mean whereas the response of the SoundLab is estimated separately, the results of which are then discarded.

It can be seen in both cases (Figures 5(a) and 5(b)) that one component-mean is very close to the expected angle of arrival while the remaining estimates are related to the SoundLab acoustic response.

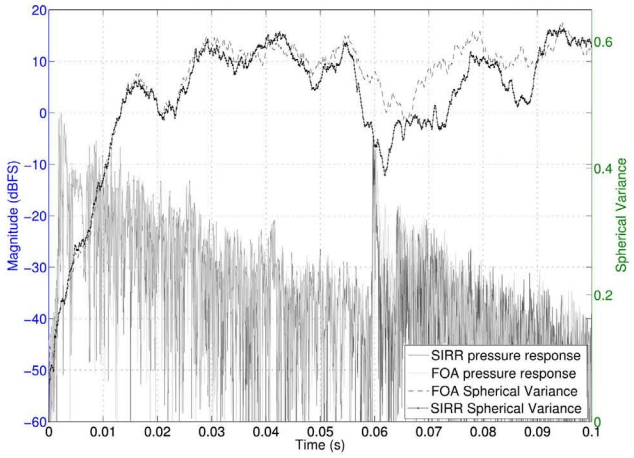
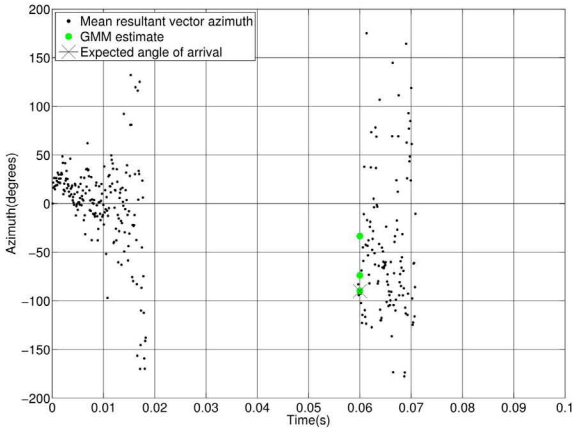
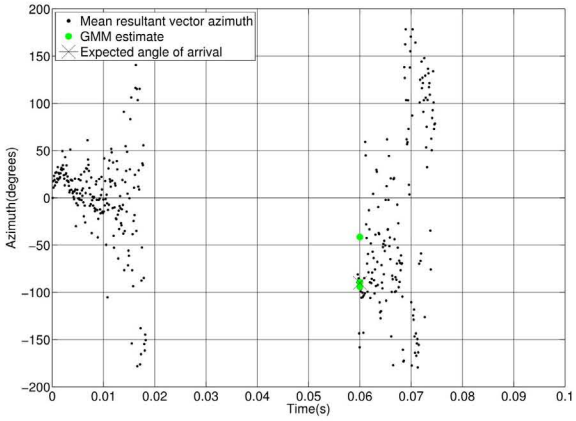


Figure 4: *Composite plot showing measured FOA and SIRR Sound Pressure Level envelopes (dBFS) and the associated spherical variance measured for each technique. This example shows the direct sound from the loudspeaker (including the early response of the SoundLab) and a single reflection occurring at* $t = 60ms$.



(a) FOA Angle



(b) SIRR angle

Figure 5: *Plots of the direction of the mean resultant vector over time for an impulse response containing a reflection at* $t = 60ms$. *Results with diffuseness* > 0.55 *have been omitted for clarity. The cross represents the expected angle of arrival* ($-90°$). *The larger dots (GMM component means) show the estimated angle of arrival*

7. LISTENING TESTS

Informal listening tests were undertaken in the SoundLab to evaluate the perceived performance of an FOA and SIRR based systems. In previous studies [1] it was found that because the musician was generating a sound that was subsequently auralised, it is highly unlikely the musician would be able to exactly repeat that sound affecting the reliability of experiments. In order to specifically compare the performance of the auralisation system it was necessary to design a repeatable experiment where the source signals remained unchanged with each repetition. Therefore, this listening test was designed as a passive test where test subjects were asked to sit in the sweet-spot of the loudspeaker array and listen to pairs of brief musical samples which were played through a directional loudspeaker positioned in front of them (imitating the musical instrument) and auralised in real-time with known impulse responses using FOA and SIRR. In this experiment, the test subject is not engaged in the physical act of playing their instrument and so is able to focus their attention solely on the audio stimuli presented to them. They will therefore be more sensitive than performing musicians to subtle differences in the presented acoustic response. Therefore this test should be viewed only as conservative comparison between the two spatialisation techniques.

The listening test was an A/B (hidden reference) test where listeners compared pairs of musical samples (where each sample had been auralised with an impulse response then rendered using FOA or SIRR) and asked to rate on a five-point Likert scale how similar or different they thought the two sounds were (1 being most similar and 5 being most dissimilar). The test pairs were short musical samples played through the loudspeaker in front of the test subject (a short cello sample playing legato (Source 1), short clarinet sample playing staccato (Source 2) and a sustained long note from a clarinet (Source 3)). This sound was auralised in real-time with an impulse response consisting of either (a) - a synthesised coherent reflection or (b) - a measured stage impulse response obtained in the Grand Hall of Glasgow City Halls or (c) - no synthesised acoustic response. The test subjects could listen to each excerpt as many times as they liked before recording their answer. The test subjects were not given any visual reference and were asked to face forward at all times. A number of null tests were introduced where both samples were the same in each pair and a single example test was presented to the listener at the beginning which was not included in the results.

There were 47 randomised combinations of stimuli in total. There were six volunteers, all between the ages of 24 and 32 (4 male, 2 female). Most of the test subjects had a background in audio engineering or acoustics, all of the volunteers had some experience of music performance. All test subjects reported no significant hearing loss.

It was expected due to the different rendering methods used, the test subject would consistently identify one method over another due to the differing ability to accurately place and reproduce the reflection. Consequently, when the test subject was asked to

rate the similarity of (a) no reflection against (b) SIRR or FOA rendered reflection, it was expected that they would report a larger difference with one technique than the other. Furthermore, it was expected that the test subjects would report significant audible differences when a musical sample, auralised with a stage acoustic impulse response, was rendered using (a) SIRR or (b) FOA techniques.

8. LISTENING TEST RESULTS

The results in figure 6 show how similar or dissimilar the participants thought the musical pairs were when one of the samples was auralised with a single reflection using FOA or SIRR and the other sample was played without a simulated reflection (i.e. direct sound only). It can be seen that in most cases, auralising with either technique produces similar median scores throughout. Furthermore, there is no discernible pattern by which one technique results in larger differences than the other.

For each reflection angle, it can be seen from the responses that the staccato clarinet sound source (source 2) resulted in reflections being identified more easily than the clarinet tone or legato cello. There is also a general indication that the presence of a reflection was more easily detected when there was a high angular separation between the reflection and the sound source.

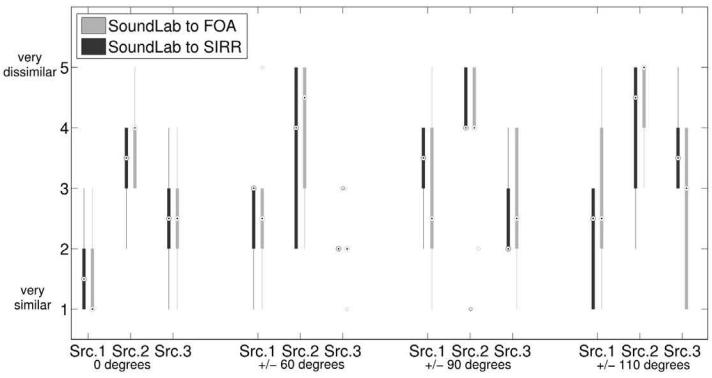


Figure 6: *Listening test results for sounds auralised with a single reflection rendered with FOA or SIRR compared to no reflection. Thick lines indicate 25 and 75 percentiles, thinner lines show the extremities of the data points, dots within boxes indicate the median while circles indicate outliers*

The results in figure 7 show how similar or dissimilar the participants thought the musical pairs were when musical samples were auralised with a stage acoustic impulse response rendered with FOA or SIRR. The results show that test subjects could discern a slight difference between the musical samples and that this difference was consistent even if the source type was altered.

9. DISCUSSION

Figure 4 showed that for a single auralised reflection, the spherical variance is lower when it is rendered using SIRR than when using FOA. This illustrates that the intensity vectors, indicating the direction of arrival of the reflection, are more tightly clustered

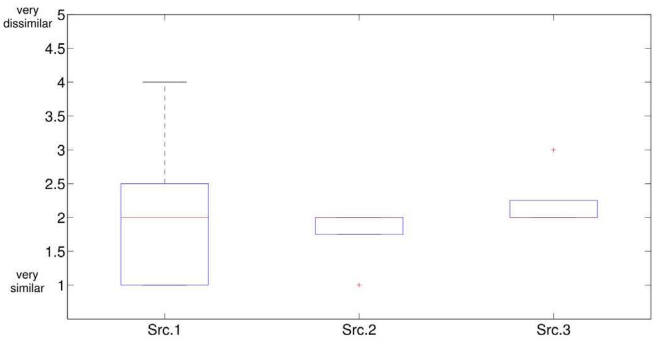


Figure 7: *Listening test results for sounds auralised with a measured impulse response rendered with FOA compared with SIRR. Central lines indicate the median response while box edges indicate 25th and 75th percentiles, outliers are indicated by crosses*

when using SIRR and therefore less ambiguous in terms of spatialisation. This is to be expected for the example shown as the reflection arrived from the same direction as one of the loudspeakers. When using SIRR, this reflection would be rendered exclusively with VBAP (due to it being non-diffuse) and therefore only one loudspeaker would be active when recreating this sound. In FOA, however all the loudspeakers are active at all times, therefore more loudspeakers would be contributing to recreate this reflection which may have contributed to a wider distribution of intensity vectors and hence a higher value of spherical variance. When single reflections were auralised from a direction between two loudspeakers, the spherical variance increased slightly when using SIRR but remained mostly unchanged when using FOA. This is consistent with operation of both spatial audio techniques where the localisation quality of FOA can be made to be largely independent of the direction of arrival. The localisation quality of VBAP will reduce slightly when the direction of arrival is between two loudspeakers separated by a large distance.

The GMM was shown to perform well when estimating the angle of arrival of reflections. It was found however that the SoundLab acoustic response introduced a noise floor into the estimation which increased as the time of arrival of the reflection decreased. This method is useful for assessing measured impulse responses but requires improvement for use in assessing auralised reflections in non-anechoic conditions. In this case, a more rigorous clustering approach could improve the accuracy of the results.

The results presented in figure 6 and figure 7 provide an initial indication that for auralisation of single reflections and measured stage acoustic impulse responses, SIRR and FOA perform equally well in the context of stage acoustic experiments although subtle differences can be discerned between the two techniques. A general trend in the listening test results indicate that in the presence of a directional sound source (representing a musical instrument), a listener may have more difficulty detecting the presence of a reflection when it arrives at a similar angle to the sound source. This is thought to be predominantly due the masking effects caused by the musical instrument. Similarly, the results also imply that detection of reflections can be affected by the type of musical expression. In this experiment, the reflections appear to have been detected more easily when the sound source was a clarinet playing staccato notes. This is thought to be due to the increased transients, coupled with

note spacing allowing the auralised reflection to be more easily detected. This however does not appear to be the case when the sound is auralised using a stage acoustic impulse response as the median results in figure 7 are very similar.

Overall, the results provide an initial indication that SIRR and FOA-based auralisation techniques perform equally well in the context of stage acoustic laboratory tests for a listener positioned in the sweetspot. However, mainly due to the low number of untrained participants there is significant uncertainty within the results, therefore further study with musician test subjects is required in order to allow for a more concrete comparison. Considering the potential advantages that SIRR offers in terms of detailed modification and analysis of impulse responses, this study provides an indication that SIRR is a viable technique for future research. This could potentially allow researchers to take advantage of the complex transformations and analysis that SIRR is capable of in addition to providing a simulation environment where the listeners impression of the soundfield is less dependent on them being located exactly in the sweetspot.

10. CONCLUSIONS

This article has demonstrated how SIRR can be used in the context of stage acoustic laboratory experiments to provide a detailed analysis of an impulse response presented to a test subject and also to re-synthesise a measured B-format impulse response using a combination of VBAP and de-correlated speaker feeds. Informal listening tests provided an initial indication that passive listeners could detect small differences between impulse responses auralised in real-time with SIRR and with FOA. Future work will focus on development of the SIRR technique for use specifically in stage acoustic laboratory experiments.

11. REFERENCES

[1] A. C. Gade, "Subjective room acoustic experiments with musicians," ISSN 0105-3027 32, Technical University of Denmark, 1982.

[2] K. Ueno, T. Kanamori, and H. Tachibana, "Experimental study on stage acoustics for ensemble performance in chamber music.," Acoust. Sci. Tech. 26, 4 (2005), 2005.

[3] J. Brereton, D. Murphy, and D. Howard, "A loudspeaker-based room acoustics simulation for real-time musical performance," Proc. of the 25th UK Audio Engineering Society conference in association with the 4th International Symposium on Ambisonics and Spherical Acoustics, 2012.

[4] J. Merimaa and V. Pulkki, "Spatial impulse response rendering," Naples, Italy, October 2004, Proceedings of the 7th International Conference on Digital Audio Effects (DAFx), pp. 139–144.

[5] A. Guthrie, S. Clapp, and N. Xiang, "Using ambisonics for stage acoustics research," Toronto, Canada, 2013, Proc. of the International Symposium on Room Acoustics.

[6] S. Tervo, T. Korhonen, and T. Lokki, "Estimation of reflections from impulse responses," Melbourne, Australia, August 2010, Proc. of the International Symposium on Room Acoustics.

[7] A. Politis, T. Pihlajamaki, and V. Pulkki, "Parametric spatial audio effects," York, UK, September 2012, Proc. of the 15th International Conference on Digital Audio Effects (DAFx-12).

[8] V. Pulkki, "Virtual sound source positioning using vector base amplitude panning," Proc. of the Audio Engineering Society, June 1997, vol. 45.

[9] T. Pihlajamaki, "Multi-resolution short-time fourier transform implementation of directional audio coding," M.S. thesis, Helsinki University of Technology, Faculty of Electronics, Communications and Automation, Department of Signal Processing and Acoustics, Espoo, Helsinki, August 2009.

[10] J. Vilkamo, "Spatial sound reproduction with frequency band processing of b-format audio signals," M.S. thesis, Helsinki University of Technology, Faculty of Electronics, Communications and Automation, Department of Signal Processing and Acoustics, Espoo, Helsinki, May 2008.

[11] A. Southern and L. Savioja, "Spatial high frequency extrapolation method for room acoustic auralization," York, UK, September 2012, Proc. of the 15th International Conference on Digital Audio Effects (DAFx-12).

[12] A. Farina, "Simultaneous measurement of impulse response and distortion with a swept-sine technique," Paris, France, 2000, Proceedings of the 108th Audio Engineering Society Conference.

Proc. of the EAA Joint Symposium on Auralization and Ambisonics, Berlin, Germany, 3-5 April 2014

DEPENDENCY OF THE FINITE-IMPULSE-RESPONSE-BASED HEAD-RELATED IMPULSE RESPONSE MODEL ON FILTER ORDER

Jian Zhang, Chundong Xu, Risheng Xia, Junfeng Li, Yonghong Yan

Institute of Acoustics, Chinese Academy of Sciences
No. 21, Beisihuan Xilu, Haidian, Beijing, 100190, China

ABSTRACT

Various approaches have been reported on HRIR modeling to lighten the high computation cost of the 3-D audio systems without sacrificing the quality of the rendered sounds. The performance of these HRIR models have been widely evaluated usually in terms of the objective estimation errors between the original measured HRIRs and the modeled HRIRs. However, it is still unclear how much these objective evaluation results match the psychoacoustic evaluations. In this research, an efficient finite-impulse-response (FIR) model is studied as a case study which is essentially based on the concept of the minimum-phase modeling technique. The accuracy dependency of this modeling approach on the order of FIR filter is examined with the objective estimation errors and the psychoacoustic tests. In the psychoacoustic tests, the MIT HRIR database are exploited and evaluated in terms of sound source localization difference and sound quality difference by comparing the synthesized stimuli with the measured HRIRs and those with the FIR models of different orders. Results indicated that the measured hundred-sample-length HRIRs can be sufficiently modeled by the low-order FIR model from the perceptual point of view, and provided the relationship between perceptual sound localization/ quality difference and the objective estimation results that should be useful for evaluating the other HRIR modeling approaches.

1. INTRODUCTION

Head-related impulse responses (HRIRs) play an important role in binaural 3-D audio rendering, which is generally realized by convolving the input stimulus with HRIRs. The direct way to obtain the HRIR from a given source location is to measure the HRIR at the ear drum for the impulse placed at the source [1]. However, the measured HRIRs are always a couple of hundred-sample lengths, which results in the high computational cost for real-time applications especially when simultaneously rendering multiple sound sources. To overcome these problems, various approaches have been reported to model HRIRs in the temporal and/or spectral domain [2][3]. The HRIR modeling approaches that have been reported in the literatures can be roughly classified into three categories: the physical-based computational approach [4][5], the parametric modeling approach [6][7], and the filter-based modeling approach [8][9].

As the simplest physical computational approach, the structural model is composed of different basic filters each of which is used to model the acoustic effects of each component of human body on wave propagation in an anechoic environment. Though this model is conceptually simple to implement, it is very difficult to estimate the model parameters from the geometrical measurements, especially for pinna [10]. Moreover, it is still unknown how to deal with the sources coming from the back half space [4]. Another well-known computational model is the boundary-element method (BEM), which provides an elegant way of the partial differential equations that describe acoustic wave propagation around a physical object [5]. The disadvantage of this approach include the difficulty in getting accurate surface meshes (especially for pinna) and the high computational cost. The negative aspect of these physical computational approaches is that they made a number of assumptions, which may remove some essential information which are necessary for producing a realistic acoustic environment simulation.

The parametric modeling approaches, which attempt to functionally represent the HRIRs, generally first model the measured HRIRs using a set of parameters that are further used for HRIR synthesis. For instance, Evans et al. suggested a form of continuous orthogonal representation in which the HRTFs were expressed as a weighted sum of surface spherical harmonics [6], and Kistler and Wightman reported to approximate the minimum-phase HRTFs with principal components analysis (PCA) [7]. But the use of such models in systems still has many draw backs. One of them is the HRIR implementation which, even if greatly compressed, requires a large computational cost to uncompress or recover the data.

The FIR and IIR filters are also widely used for HRIR modeling in the temporal and/or spectral domain. Bolmmer and Wakefield presented to design the IIR filter based on the error criteria of log-magnitude spectrum differences [9]. Asano et al. investigated the abilities of IIR filters with different orders for modeling individual HRIRs, and showed that a 40th-order IIR filter yielded good approximation of individual HRIRs in terms of sound localization difference, with the exception of increased front-back confusions in frontal incident angles [11]. Though IIR is able to model HRIRs with a quite low order, it is very difficult for IIR filter to be interpolated between discrete positions.

The performance of most HRIR models have been widely evaluated usually in terms of the objective estimation errors between the original measured HRIRs and the modeled HRIRs. However, it is still unclear how much these objective evaluation results match the psychoacoustic evaluations. In this paper, we focus on a FIR approach for HRIR modeling on the concept of the minimum-phase approximation technique. Main attention was paid to investigate the accuracy dependence of this FIR-based modeling approach on the order of FIR filters through psychoacoustic tests in terms of sound source localization difference and sound quality difference. And one objective evaluation errors was used to investigate the effectiveness of this FIR models. Psychoacoustic test results demonstrated that the measured HRIRs can be sufficiently modeled by the low-order FIR model. And the relationship

between perceptual sound localization/ quality difference and the objective estimation results is useful for evaluating the other HRIR modeling approaches.

2. METHOD

The implementation of the filter-based modeling approaches can be in the time or frequency domain. The modeling approach studied here is in the frequency domain. More specifically, the frequency responses of the measured HRIRs are approximated on the minimum phase theory, which consists of the following steps (We use the front HRIR (0°, 0°) of MIT database as an example):

1. *ITD estimation of HRIRs.*

   The commonly used approach for ITD estimation is based on the cross-correlation between left and right channels of HRIRs [7]. The ITD is calculated in this work through estimate the time delay of each HRIR, which is determined as the time at which the HRIR becomes non-zero using the threshold-based detection technique, and yielded the almost identical results with cross-correlation method [2]. As shown in the figure 1, the time delay is 33 samples (the red part).

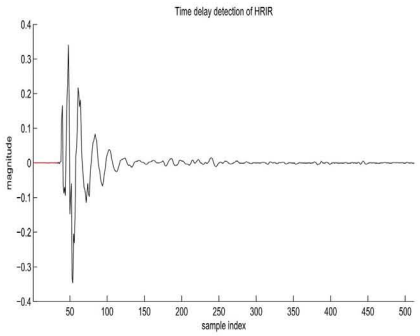


Figure 1: Example of time delay detection of HRIR.

2. *Amplitude approximation of the minimum-phase HRIRs with the different FIRs of different orders.*

   The minimum-phase HRIRs, h'(t), are first derived by removing the zeros at the beginning of HRIRs based on the estimated ITDs. The corresponding transfer functions are denoted as H'(k) in the frequency domain. Then the amplitude responses of H'(k) are approximated by the FIR filters. The coefficients of the FIR filters to be estimated are eventually determined by using a linear least square error function [12]. Note that given different FIR orders, different coefficients of FIR filters can be derived. In other words, the FIR models with different orders should yield different abilities in approximating the amplitude response H'(k) of the minimum-phase HRIRs. Figure 2 depicts the frequency response of minimum-phase HRIR and the frequency response of FIRs with the order of 47 and 68.

3. *Modeled HRIRs synthesis by adding ITD cues.*

   The time domain modeled HRIRs h(t) can be obtained using the FIR filters, followed by supplementing the ITD cues. Actually, it is needn't add the zeros to the FIR directly. This procedure is essentially implemented by adding the zeros to the output signal when the convolving of input signal and

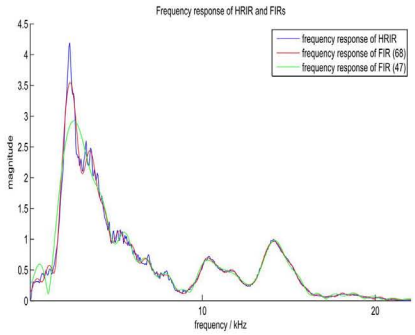


Figure 2: Frequency response of HRIR and FIR.

FIR is done. So the number of zeros will not affect the efficiency of this model. Figure 3 shows the original HRIR and the FIR with the order of 47. (For easily compare, the start point of FIR in the figure was moved 33 samples.)

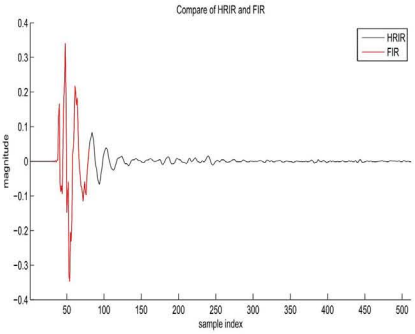


Figure 3: Compare of original HRIR and modeled FIR.

3. PERCEPTUAL EVALUATIONS

3.1. Data

The HRIRs were provided by MIT measured using a KEMAR dummy head [1]. Gardner and Martin made an assumption that the dummy head is perfectly symmetrical, so the HRIRs need be collected for only one ear. This assumption allowed them to mount two different pinnae on the KEMAR, and the HRIRs associated with both pinna types could be collected simultaneously. The HRIRs used in this paper are measured with the right ear.

To investigate the dependence of the ability of the considered HRIR modeling approach on the order of FIR filters and find out the minimum order of the FIR for modeling the measured HRIRs, various FIR orders were designed according to the E series standard [13]. Since the length of the measured HRIRs were 512 samples, the examined FIR orders were eventually determined as 10, 15, 22, 33, 47, 68, 91, 121, 178, 200, 261, 383, 464. Given a FIR order, the modeled HRIRs can be computed using the approach described in the previous section. In our tests, the directions of horizontal plane and vertical plane are evaluated. The horizontal angles of the HRIRs are from -90° to 90° at the interval of 30°, and the vertical angles were from -30° to 90°, the interval is 30°.

To generate the stimuli for psychoacoustic tests, three signals were exploited, including one broad-band white noise, one male

speech signal and one telephone ring signal, with the length of 2s, 3s and 2s respectively. The stimuli were eventually generated by convolving three types of signals with the HRIRs modeled with the FIRs of different orders and adding time delay. The generated stimuli were subsequently presented to subjects for psychoacoustic tests.

3.2. Subjective evaluation

A total of 10 subjects (5 male and 5 female) with normal hearing were recruited and paid for their participation in psychoacoustic tests. The subjects were aged from 23 to 28. In tests, stimuli were presented to the subjects at a comfortable listening level with Sennheiser HD 280 Pro headphones. Two psychoacoustic tests were carried out for comparing the stimuli generated by the modeled HRIRs and those by the measured HRIRs, in terms of sound quality difference and sound localization difference, respectively. In each evaluation (for quality or localization difference), each subject listened to a total of 1512 stimuli (3 signals $\times$ 12 DOAs $\times$ (13 orders + 1 original) $\times$ 3 times), where each stimuli were presented three times. The stimuli were grouped into 108 sets in each of which 14 stimuli (processed by the FIR filters with 13 different ordered and the measured original HRIRs) were randomly presented to subjects. In listening tests, the 108 sets were further divided into 6 sessions with 18 sets in each session. The subjects were asked to have a break after one session.

For the perceptual tests, the paired comparison evaluation was used, in which one stimulus was generated by convolving the dry signal with the measured HRIR and the other with the FIRs of different orders. The presentation order of the stimuli sets and the stimuli in each set were randomized for each subject. The stimuli could be listened to several times until the subject made a decision. For each paired comparison, subjects provided a score on the five-grade scale based on his/her preference in terms of the degree of difference in sound quality or sound localization. The detail specification of the five-grade scale is shown in Table. 1. During the test, no feedback information was given to the subjects.

Table 1: Five-grade score scale using in the psychoacoustic listening tests and its description.

Score	Description
1	Exactly different
2	Different
3	Uncertain
4	Almost same
5	Exactly same

3.3. Objective evaluation

In order to objectively evaluate the model against the original measured HRTF, and further find the relationship between the objective and subjective evaluations, the spectral distortion (SD) was considered as error measure [10].

$$SD = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(20\log_{10}\frac{|H(f_i)|}{|\tilde{H}(f_i)|}\right)^2}\ [dB] \qquad (1)$$

where H is the frequency response of original HRIR, $\tilde{H}$ is the frequency response of modeled HRIR (FIR), and N is the number of available frequencies in the considered range, that limited between 500 Hz and 16 kHz. For calculate $\tilde{H}$, the zeros are added to the FIRs.

3.4. Results

The overall psychoacoustic results in terms of sound quality difference and localization difference are plotted in figure 4. As shown in figure 4, a FIR filter with 68 coefficients (Note that the zeros that represent the time delays are not considered) is sufficient in both sound quality and localization. That is, the FIRs with 68 coefficients (not 512 samples as in the measured HRIRs) are able to provide the quite similar perceptual sensation. The FIR-based HRIR modeling approach greatly reduced the length of HRIRs, which further reduced the computational cost and speed up the synthesis procedure of binaural signals. Furthermore, if only sound source localization performance is considered as evaluation criterion, the minimum order of FIRs can be further decreased to 47, at which the sound quality will be slightly different from that by the measured HRIRs.

As the three signals used in our psychoacoustic tests exhibited different energy distribution in the time-frequency domain, the dependency of the modeling ability of FIR-based approach on the FIR order was further investigated by looking at its relationship with the stimuli type. The perceptual results in terms of sound quality difference and sound source localization difference are plotted in figure 5 and figure 6, respectively. From these results, it is noted that the minimum order of FIR that provided the acceptable sound quality and sound localization is dependent on the stimuli type. For example, the FIR order that yielded sufficiently acceptable sound quality and localization for the telephone ring signal is lower than that for the male speech and white noise signals. This difference might come from the difference in energy distribution of each stimulus in the time-frequency domain.

Figure 7 depicts the SD of FIRs with different orders. As shown in figure 7, the SD is nearly 2 dB where the order of FIR is 47, which obtain acceptable source localization performance. If the sound quality is concerned, which means the order of FIR would be 68, the SD is nearly 1 dB. When the SD is larger than 4, correspond to the FIR order is small than 22, both the sound quality and source localization performances are poor.

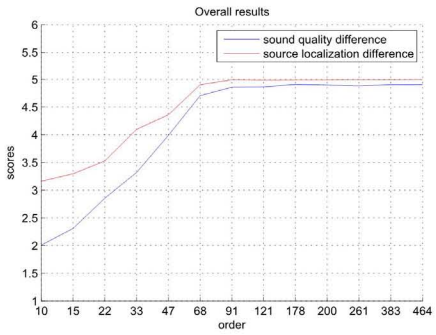


Figure 4: Overall results.

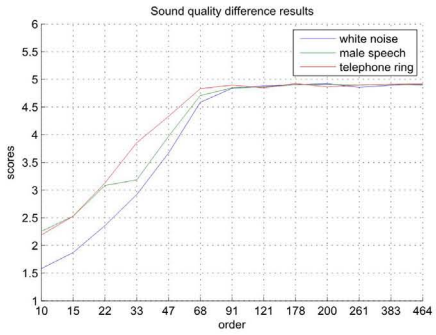


Figure 5: Sound quality difference results of different stimuli.

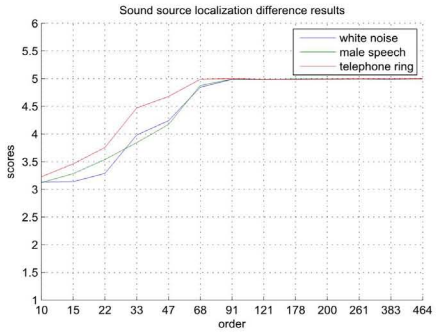


Figure 6: Sound source localization difference results of different stimuli.

4. CONCLUSION

In this paper, a FIR-based HRIR modeling approach based on minimum phase theory was studied. In this paper, the HRIR is approximated by a FIR filter and ITD cues. The main attention in this paper was paid to the performance dependence of this modeling approach on the order of FIRs through psychoacoustic tests in terms of sound quality difference and sound source localization difference. Psychoacoustic test results indicated that the measured HRIRs with the length of hundred samples can be perceptually modeled by the low-order FIRs with a dozen of coefficients. And the performance is further evaluated by objective quantity, the relationship of the objective and subjective evaluation would be helpful for other HRIR modeling methods.

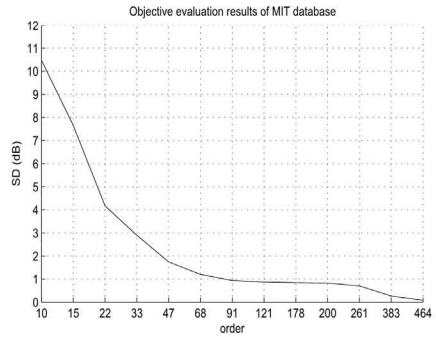


Figure 7: objective evaluation results of MIT database.

Acknowledgment

This work is partially supported by the National Natural Science Foundation of China (Nos. 10925419, 90920302, 61072124, 11074275, 11161140319, 91120001), the Strategic Priority Research Program of the Chinese Academy of Sciences (Grant Nos. XDA06030100, XDA06030500) and the National 863 Program (No. 2012AA012503).

5. REFERENCES

[1] W. G. Gardner and K. D. Martin, "Hrtf measurements of a kemar," *J. Acoust. Soc. Am.*, vol. 97, no. 6, pp. 3907–3908, 1997.

[2] J. Sandvad and D. Hammershoi, "Binaural auralization, comparison of fir and iir filter representation of hirs," in *Audio Engineering Society Convention 96*, Feb 1994.

[3] J. Huopaniemi and M. Karjalainen, "Review of digital filter design and implementation methods for 3-d sound," in *Audio Engineering Society Convention 102*, Mar 1997.

[4] C. P. Brown and R. O. Duda, "A structural model for binaural sound synthesis," *IEEE Trans. on Speech and Audio Processing*, vol. 6, no. 5, pp. 476–488, 1998.

[5] B. F. G. Katz, "Boundary element method calculation of individual head-related transfer function, impedance effects and comparisons to real measurements," *J. Acoust. Soc. Am.*, vol. 110, no. 5, pp. 2449–2455, 2001.

[6] M. J. Evans, J. A. S. Angus, and A. I. Tew, "Analyzing head-related transfer function measurements using surface spherical harmonics," *J. Acoust. Soc. Am.*, vol. 104, no. 4, pp. 2400–2411, 1998.

[7] D. J. Kistler and F. L. Wightman, "A model of head-related transfer functions based on principal components analysis and minimum-phase reconstruction," *J. Acoust. Soc. Am.*, vol. 91, pp. 1637–1647, 1992.

[8] A. Kulkarni and H. S. Colburn, "Efficient finite-impulse-resonse filter models of the head-related transfer function," *J. Acoust. Soc. Am.*, vol. 97, no. 5, pp. 3278, 1995.

[9] M. A. Blommer and G. H. Wakefield, "On the design of pole-zero approximations using a logarithmic error measure," *IEEE Trans. on Speech and Audio Processing*, vol. 42, no. 11, pp. 3245–3248, 1994.

[10] S. Spagnol, M. Geronazzo, and F. Avanzini, "On the relation between pinna reflection patterns and head-related transfer function features," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 21, no. 3, pp. 508–519, 2013.

[11] F. Asano, Y. Suzuki, and T. Sone, "Role of spectral cues in median plane localization," *J. Acoust. Soc. Am.*, vol. 88, no. 1, pp. 159–168, 1990.

[12] E. C. Levi, "Complex-curve fitting," *IRE Trans. on Automatic Control*, vol. AC-4, pp. 37–44, 1959.

[13] IEC, "Preferred number series for resistors and capacitors," in *Standard IEC 60063, International Electro technical Commission*, 1963.

[14] J. Chen, B. D. V. Veen, and K. E. Hecox, "A spatial feature extraction and regularization model for the head-related transfer function," *J. Acoust. Soc. Am.*, vol. 97, no. 1, pp. 439–452, 1995.

AURALIZATION OF SEVERAL CHURCHES AND LISTENING COMPARISON USING MULTIDIMENSIONAL SCALING APPROACH.

Pawel Malecki

[1]Department of Mechanics and Vibroacoustics,
AGH – University of Science and Technology
Krakow, Poland
pawel.malecki@agh.edu.pl

ABSTRACT

Modern auralization techniques allow to make better assessment of particular interior type for different destination and purposes. The quality and reality of acoustic recording and reproduction systems increase so the results of this kind of research are much more reliable.
The article shows the psychoacoustic comparison of different reverberant interiors. The auralization was provided using 1st order ambisonics spatial impulse responses convoluted with anechoic choral music. Listening tests were conducted within the 16-channel sound system. The subjects were tested using the pair comparison method and the results were analyzed with the multidimensional scaling approach.

1. INTRODUCTION

Although there is an increase of popularity and diversity of modern auralization techniques, they are not very often employed in a room acoustics validation and evaluation. Early works on that subject were conducted using few channels sound systems with many simplifications [1], [2]. Modern ambisonics techniques [3], [4], sound-field synthesis [5] and other complex systems [6], [7] are well known and commonly exploited but hardly to previously mentioned goals. Actual work focuses on that subject with respect to orthodox churches, so on interiors, where the acoustics features are extremely important and connected with a sense of characteristic space. The chosen orthodox churches from the southern and eastern parts of Poland were examined in the context of the orthodox choral music. The Orthodox church (referred as a building) is inseparably connected with spirituality and philosophy of the eastern Christianity. Its architecture is based on orthodox liturgy and implicates a natural cultural landscape of the east part of Europe. The acoustic measurements in orthodox churches were marginalized by now and only a few buildings were examined.
The first order ambisonics is employed in a measurement process and in listening experiments. The article shows the research methodology and the main statistically developed results.

2. MEASUREMENTS OF SPATIAL IMPULSE RESPONSES IN THE CHOSEN ORTHODOX CHURCHES IN POLAND

Measurements of SIRs (Spatial Impulse Responses) were conducted in eleven chosen orthodox churches in Poland from June to October 2012.

The measurements and the overview of basic features were provided for the churches with different architecture style, size and interior design. In order to measure SIRs, EASERA Pro software was employed as well as authors algorithms implemented with the Matlab computing language. The ISO3382 standard [8] was followed during the measurements as possible. Due to the specific goals of the research, some simplifications were made. As it was mentioned before, first order ambisonics microphone - Soundfield ST350 was used, and instead of an omni-directional source, an active loudspeaker JBL EON 515 was employed. Because of the main destination of SIRs (listening experiments), the number of measurement locations was also limited. The main placement of the measurement instruments was common for all the churches. The sound source was placed in front of the iconostasis and the microphone behind a tetrapod. This is a typical place where a priest, a deacon sings or the choral music concerts take place. Behind the tetrapod (a small table about six meters in front of the iconostasis) there are the best first seats for the audience. On the basis of the B-format pressure component, the room acoustic parameters were also calculated and some of them are listed in table 1. Except for the previously explained setup, the measuring points were located at characteristic places for orthodox churches:

- in a central point of the church (at least at two different positions),
- under the main dome,
- under the gallery,
- at the back of the church,
- at a side naves (at least three different positions),
- at other positions if the church shape suggested this.

Except for the small local wooden churches at Zdynia and Konieczna, all of the churches were made of bricks. The most Latin style characterizes the medium size Hole Trinity Church in Sanok. The churches in Wlodawa and Hrubieszow are the examples of the Byzantine-Russian style. The church in Tomaszow Lubelski is also a medium size as well as a St. Peter and Paul in Siemietycze, both in the Neo-Russian style. The churches in Bialystok, Hajnowka and the Resurrection Church in Siemiatycze are the largest ones.. The church in Hajnowka is well known from the International Festival *The Hajnowka's Orthodox Church Music Days* were best choirs from all over the world perform. The church of Divine Wisdom in Bialystok is interesting due to the fact that it is a miniaturized copy (1:3) of Hagia Sophia in Istanbul.

Table 1: *Acoustic parameters of church SIRs used for listening tests.*

No.	Church of the…\ Parameter	RT [s]	C_{80} [dB]	C_{50} [dB]	t_s [ms]	LEF	STI	G	Vol. [m^3]
1	Holy Trinity in **Sanok**	1,83	4,2	2,3	65	0,47	0,58	7,3	1650
2	Sacred Virgin Mary in **Włodawa**	2,09	2,6	0,8	93	0,57	0,58	6,6	2400
3	Assumption in **Hrubieszow**	1,82	4,6	2,5	59	0,21	0,61	3,3	1700
4	St. Nicholas in **Tomaszow Lubelski**	2,87	3,4	1,7	105	0,25	0,58	3,8	3050
5	St. Peter and Paul in **Siemiatycze**	2,2	0,7	-1,2	109	0,76	0,51	7,7	1500
6	Holy Spirit in **Bialystok**	6,53	0,2	-0,4	235	0,22	0,49	6,6	9500
7	Holy Trinity in **Hajniwka**	4,59	3,9	3,0	106	0,1	0,56	3,7	6050
8	Resurrection in **Siemiatycze**	4,74	3,9	3,0	106	0,14	0,56	5,8	6400
9	Divine Wisdom in Bialystok (**Hagia**)	6,14	0,9	-0,7	198	0,27	0,46	5,7	5950
10	Protection of the Mother of God in **Zdynia**	1,07	6,9	4,1	41	0,51	0,68	4,7	750
11	St. Basil in **Konieczna**	1,17	7,7	5,4	37	0,46	0,68	5,2	900

3. LISTENING SETUP

For the listening test, the 16-channel setup was prepared using the RME converters and Genelec 6010 monitors. The selected loudspeakers are quite small but their sensitive is 93 dB SPL with flat frequency response from 74 Hz to 18 kHz (±2,5 dB). The loudspeakers are spherically placed around the listener, whose configuration is shown in Fig. 1. This setup was installed at the anechoic room of AGH-UST using microphone stands with sphere diameter of 3,2 meters.

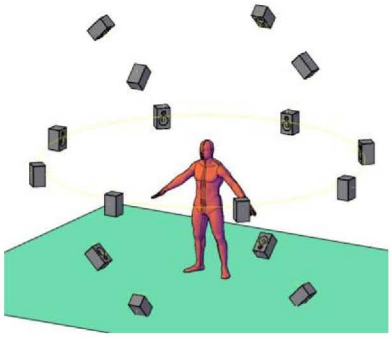


Figure 1: *The idea scheme of loudspeakers placed on sphere configuration*

The system was positioned with a laser angular meters and then calibrated and phase checked.

Figure 2: *Listening stand overview photo*

Loudspeakers of the system are situated on a sphere around a listener in three planes – above, under and on the height of listener's head. Exact angular coordination in respect to the listener are presented in the table 2.

Table 2: *Loudspeakers angular coordination in respect to the listener position*

Loudspeaker no.	φ [°]	θ [°]
1	0	0
2	45	0
3	90	0
4	135	0
5	180	0
6	-135	0
7	-90	0
8	-45	0
9	68	-45
10	158	-45
11	-113	-45
12	-23	-45
13	23	45
14	113	45
15	-158	45
16	-68	45

4. STIMULI PREPARATION

The human voice is often regarded as one of the most beautiful musical instruments. When the human voice is multiplied (like in choir), as the result of this mixture there is a sound of a very rich and interesting timber. Choral music is inseparable from Orthodox church and its tradition so as a samples in listening test, choral music exerts were used.

4.1. Raw sounds recording

In order to obtain raw audio material (without any reflections or other type of response), choral free field recording were made. Recording of chamber orthodox choir (8 singers) was performed at anechoic room of AGH-UST. The whole choir was recorded using multitrack setup using one type of microphones (Rode NT5 with cardioid capsule) for each voice. Soprano, alto, tenor and bass tracks were separately recorded and prepared for next re-

search stages. Recorded music samples were in moderate tempo and not complex harmony, with neutral, and universal lyrics (*halleluiah*).

4.2. Raw audio samples convolution with SIRs

Finally, obtained SIRs were convoluted with a raw choir music samples. In order to place each voice in the right position of sound plan, each convolution was performed for every voice separately. As a result there were 176 convolutions (11 churches x 4 ambisonics components x 4 choir voices). After that, the B-format tracks were converted to 16-channel setup, acording to [3] and the coordinates shown on table 2.
Choir voices placement were performed using VBAP (Vector Based Amplitude Panning) technique [9]. Voices (phantom images) were placed typically for regular SATB choir accordingly to data in tables 2 and 3. Angular positions of particular voices in table 2 are the trigonometric result of typical distances between singers and 6.5 m distance from listener position. Table 3 shows calculated gain factors for specific loudspeakers of listening setup, based on VBAP.

Table 3: *Gain factors of specific loudspeakers of setup*

Voice	φ [°]	Loudspeaker no.	Gain factor
Soprano	-20	1	0,78
		8	0,63
Alto	-7	1	0,98
		8	0,19
Tenor	7	1	0,98
		2	0,19
Bas	20	1	0,78
		2	0,63

5. LISTENING TEST

The main purpose of the performed psychoacoustic procedure is to ranking measured churches in sense of *acoustic esthetics*. It is difficult to precise this kind of criteria. So, in order to find out relations between the listener preferences and the physical features of the rooms, the MDS (Multidimensional Scaling) method was employed [10]. Obtaining data about objects similarity and preferences for MDS analysis was performer using paired comparison. For 11 churches, there are 55 unique possible pairs to compare. Each test round was designed using 2IFC procedure (2 stimuli, intervaled forced-choice). After two choral music excerpts of length about 15 s each, *located* in different churches, the listener's task was to choice which of them sounds better and how much do they differ. To clarify and unify listeners task, the exact question was stated before each survey:

- *How similar are the following sound excerpts representing different interiors, where 0 denotes to totally different and 100 refers as identical?* {similarity test}
- *In which of presented auralized interiors would you prefer to listen to this kind of music? In other words: Which of the presented sounds you like more?* {preference test}

Listeners answered to these questions after each round of two sounds using tablet touch interface. Fig. 3 shows the application form window used during the tests.

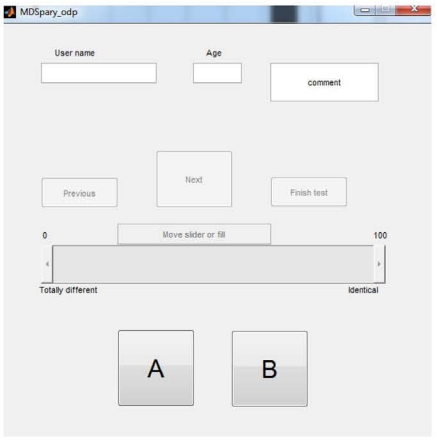


Figure 3: *Form window used during the tests*

Actual tests were followed by a few training rounds to introduce test subjects with exemplary stimuli and to practice whole test procedure. Test subjects were placed in the center of loudspeaker sphere individually to achieve similar sweet spot condition for every listener. During the tests, only the subject and the operator remained in the anechoic chamber. Whole test last for about 40 minutes so a short break in the middle of it was obligatory. Every test round was introduced by a lector, and the listeners had the opportunity to repeat every round if needed. Listening group counted of 20 people from 20 to 35 years old within 6 female. Listeners had different listening or acoustics experience and different confession. None of the listeners had impaired auditory system.

6. STATISTICAL SIGNIFICANCE OF RESULTS

Actual work discusses results of a preference test only. First step of statistical analysis is the correlation between personal listeners results. Average value of Pearson's coefficient equals 0,38 for n = 20. Zero hypothesis which assumes no correlation between answers of listeners was not rejected in 40 cases on 180 possible pairs, at 0,95 significance level. The conclusion is that alternative hypothesis assuming significant correlation of the particular results cannot be accepted. For more intuitive and visual correlation analysis, metric MDS [10] is used to visualize correlation matrix (fig. 4). It could be noticed that there are two clusters of subjects relatively close to each other. This suggests to perform analysis with respect to noticed division – separately for both groups, marked in fig. 4 as group A and B.

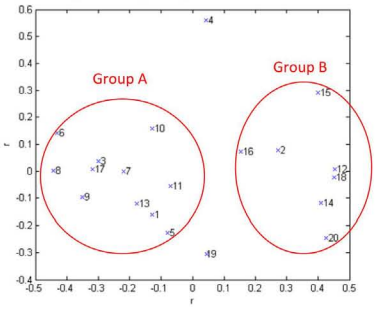


Fig 4: *Graphical interpretation of correlation matrix of results of particular subjects*

7. RESULTS

Collected data was processed in two ways. Fig. 5 shows preferences test results using Thurstone-Mosteller's least square approach [11]. Presented results are based on the probability analysis of listeners preferences in paired comparison test. Mentioned probability is calculated on the base of quantity when object *i*, was marked as preferred accordingly to other object *j*. This kind of scaling is performed for both statistically significant groups of subjects.

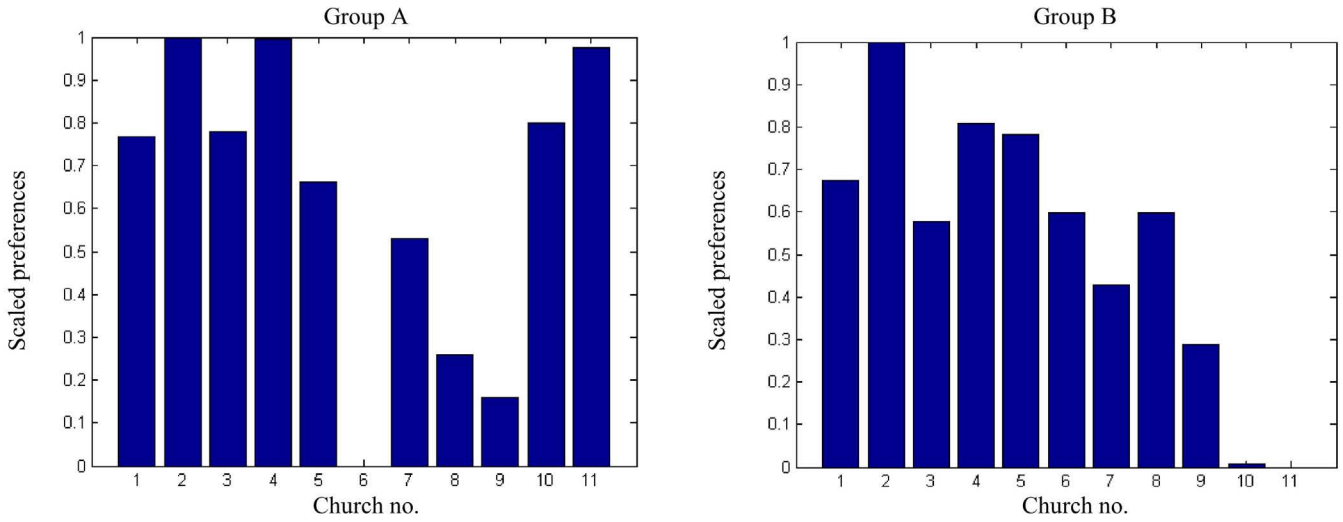


Fig 5: *Listeners preferences calculated using Thurstone-Mosteller's least square approach of data obtained in paired comparison test*

After comparison of diagrams in fig. 5 it could be noticed that the main difference between considered groups is for objects 6-11. Group A prefers objects 10 and 11 much more than objects 6, 8 and 9. Opposite tendencies can be noticed for group B. Both groups prefers objects 2 and 4 the most.

Second approach to calculation is based on non-metric MDS scaling. Difference between objects is calculated on the basis of probability of preferences according to transformation in [10]. Graphical interpretation of preferences for both statistical groups are shown in fig. 6. Applying non-metric MDS scaling within 2 dimensions resulted with stress factors of 0,11 for both groups. For random data and the same analysis parameters stress factor would equals 0,21. So according to [10], actual fitting is rated as good.

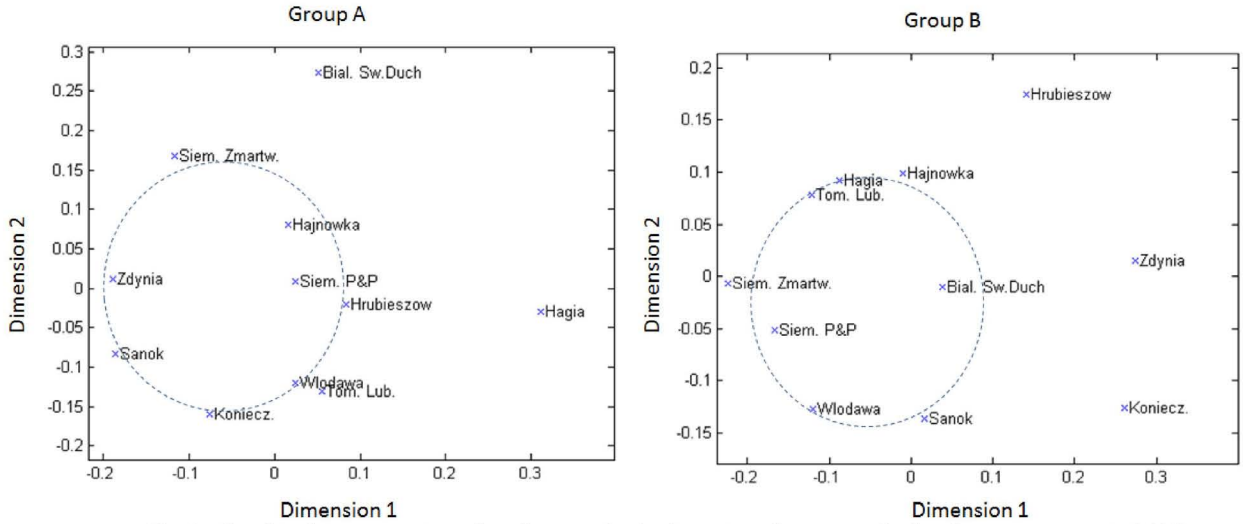


Fig 6. *Graphical interpretation of preferences for both statistical groups calculated using non-metric MDS*

For both diagrams in fig 6 there are two geometrically arranged aggregations of objects and some separated ones. For group A this separated objects are very big and reverberant and for group B very small and acoustically dead compared to others. On the basis of fig 5 and 6 it can be concluded that noticed statistical groups differs mostly in sense of negative preference. Analysis of individual subjects does not suggest any special features of any groups (age, acoustic-music experience, sex etc.) Smaller group

B contains three orthodox listeners versus one orthodox in a group A. This clue seems to be not strong enough to make some general conclusion about it.

8. CONCLUSIONS

Measurements of 11 SIRs in several orthodox churches were conducted as well as listening test using auralized samples of choral music. Tests were conducted in order to find listeners preferences about highly reverberant rooms in aspect of choral music. Auralization was performed using ambisonics and VBAP technologies. First order ambisonics was employed at every stage of research. Conducted acoustic measurements in orthodox churches in central Europe on that scale are one of the first.
Listening experiments were performed using psychoacoustic methodology and modern statistical MDS approach. Correlation analysis of results revealed two different groups of listeners. They had different preferences especially about extreme reverberant conditions. They had completely opposite opinion about very long and short reverberation time in context of choral orthodox music. Based on conducted research very interested conclusion can be stated that listeners were divided not with aspect to *what do they like*, but rather in *what do they not like*. All listeners rated medium size churches with big central dome as most preferred. Future research on current subject is planned with employment of higher order ambisonics and different approach in on place SIR measurement in order to achieve even more realistic auralization.

9. REFERENCES

[1] Y. Ando, "Subjective preference in relation to objective parameters of music sound fields with a single echo", *J. Acoust. Soc. Am*, vol. 62, pp. 1436–1441, 1977.

[2] M. Barron, "The subjective effects of first reflections in concert halls – The need for lateral reflections", *J. Sound Vibr*, vol. 15, pp. 475-494, 1971.

[3] A. Farina, R. Glasgal, E. Armelloni and A. Torger, "Ambiophonic Principles for the Recording and Reproduction of Surround Sound for Music", *Proceedings of the 19th AES Conference on Surround Sound, Techniques, Technology and Perception*, *Schloss Elmau, Germany*, 2001.

[4] C. A. Dimoulas, "Sound source localization and B-format enhancement using soundfield microphone sets", *Proceedings of the 122th AES Convention*, *Vienna, Austria,* 2007.

[5] S. Spors, H. Wierstorf, A. Raake, F. Melchior, M. Frank and F. Zotter, "Spatial Sound With Loudspeakers and Its Perception: A Review of the Current State", *Proceedings of the IEEE*, vol.101, no.9, pp. 1920 - 1938, 2013.

[6] V. Pulkki, "Spatial Sound Reproduction with Directional Audio Coding", *J. Audio Eng. Soc*., Vol. 55, pp. 503-516, 2007.

[7] W. Woszczyk, D. Ko and B. Leonard, „Convolution-based virtual concert hall acoustics using aural segmentation and selection of multichannel impulse responses", *The Proceedings of INTER-NOISE 2009, Ottawa, Canada*, 2009.

[8] ISO 3382, *Acoustics – Measurement of the Reverberation Time of Rooms with Reference to Other Acoustical Parameters*. Geneva, ISO, 1997.

[9] U. Zolzer et al., *DAFX: Digital Audio Effects*, John Wiley & Sons Ltd, 2011.

[10] I. Borg and P. J. F. Groenen, *Modern Multidimensional Scaling*. Springer, 2005.

[11] K. Tsukida and M. R. Gupta, *How to Analyze Paired Comparison Data*. University of Washington, 2011.

Proc. of the EAA Joint Symposium on Auralization and Ambisonics, Berlin, Germany, 3-5 April 2014

A REAL-TIME AURALIZATION PLUGIN FOR ARCHITECTURAL DESIGN AND EDUCATION

Lukas Aspöck, Sönke Pelzer, Frank Wefers, Michael Vorländer

Institute of Technical Acoustics,
RWTH Aachen University,
Aachen, Germany
{las,spe,fwe,mvo}@akustik.rwth-aachen.de

ABSTRACT

The role of acoustics in architectural planning processes is often neglected if the designer lacks necessary experience in acoustics. Even if an acoustic consultant is involved he might be presented with limited options after the initial planning process. Some disadvantageous decisions might be hard to reverse then. To improve and facilitate the construction process permanent immediate feedback should be given to the designer. Planning cannot be imaged today without live 3D visual rendering. But also acoustics should be rendered in real-time to provide the same type of intuitive feedback. Therefore a real-time room acoustics auralization was implemented into a popular CAD-Modeling tool. Binaural room impulse responses are continuously updated using image sources and ray tracing algorithms and convolved in real-time with audio feed from recorded sounds or the user's microphone. The CAD model can be freely modified during the simulations including geometry, surface materials and source and receiver positions. Using streaming low-latency convolution, an immediate feedback is provided to the user.

1. INTRODUCTION

Nowadays the most important tool during the design process of an architect is the CAD-editor. It enables the architect not only to use it as a planning tool, but it also offers the possibility to experience his ideas in three dimensions without erecting the physical structure. Especially when it comes to buildings and rooms which are used for the presentation of acoustical signals, such as concert halls or conference rooms, it becomes important to also include the acoustical characteristics in the design process. For the visual appearance, it is possible to use rendering tools producing a photo-realistic image of the model. To achieve the same level of quality for the audio feedback, an auralization based on room acoustics simulations is required. Room acoustic simulation algorithms are usually based on the assumption that a sound wave is interpreted as a ray which behaves in a similar way as a light ray (Geometrical Acoustics). Often a hybrid model is used to simulate a room impulse response, combining the image sources methods[1] and the ray tracing algorithm [2].

Based on the geometry as well as on surface properties, acoustic simulation software (e. g. CATT-acoustic [3]) is able to provide realistic auralizations of the room as well as a reliable prediction of the room acoustic parameters. The same simulation models are also integrated in immersive Virtual Reality systems[4], allowing the user to interact intuitively in a virtual environment while experiencing multimodal feedback. Similar room acoustics simulation techniques are expected to be applied in the entertainment industry soon, e. g. for real-time sound rendering in computer games [5]. These models however focus not on providing the realistic acoustical reproduction of the environment, but on rendering plausible audio feedback with reduced computational effort, without interfering the high computational demands of the visual rendering.

However, all of these listed applications of room acoustic simulations do not represent a suitable easy-to-use tool, which the architect can use to analyse and experience the room acoustics in his building. The available room acoustic tools are often expert tools, requiring extensive knowledge about room acoustics and the simulation techniques. Virtual Reality systems are very expensive and also do not represent a convenient tool due to the system's complexity. By integrating an auralization software into the popular 3D modeling software *SketchUp*, we are able to provide an easy to operate tool, which helps to understand the effects of room acoustics as well as enriches the architectural design process.

2. STATE OF THE ART

Extensive research in the areas of room acoustics, binaural hearing and spatial reproduction made it possible to develop algorithms to simulate [6] and auralize [7][8] virtual sound fields. Because of the steady increase of computational power, today's room acoustic simulation software (e. g. ODEON [6], CATT or EASE) is able to calculate accurate results by generating room impulse responses in a reasonable amount of time, including effects such as absorption, scattering and diffraction[9]. In various surveys, the accurarcy of different room acoustic simulation software has been validated and compared to measurements[10][11][12]. These tools help the user to predict and understand the room acoustics in the room of interest. A binaural synthesis based on measured Head-Related-Transfer-Functions[13] offers the possibility to listen to virtual sources in the room. These auralizations are however often only possible for a static receiver-source situation in the room or, in case of dynamic situations, depend on precalculated and interpolated simulation results [14] or on the usage of computer clusters[15]. Although editing options for CAD models are given in most simulation tools, the possiblities are usually very limited and an external dedicated 3D design tool is much more powerful and often preferred. To eliminate this lack of interactivity with the model, this work presents the implementation of a real-time auralization software, directly integrated in the *SketchUp* modeling software and providing binaural feedback through headphones to the user.

3. REAL-TIME AURALIZATION IN SKETCHUP

The real-time auralization tool can be seperated in three parts: the *SketchUp* plug-in, the simulation client(s) and the real-time convolution module. The concepts and aspects of their implementation of these software modules are presented in the next sections. Fig. 1 gives an overview of the system. The plug-in (Section 3.1) not only includes the graphical user interface, but also acts as a server for the scene data. It contains a network interface which provides the relevant data for the room acoustic simulation to one or more simulation clients (Section 3.2). The results of the simulation (e. g. a binaural room impulse response) are sent to the convolution module (Section 3.3) and played back to the user.

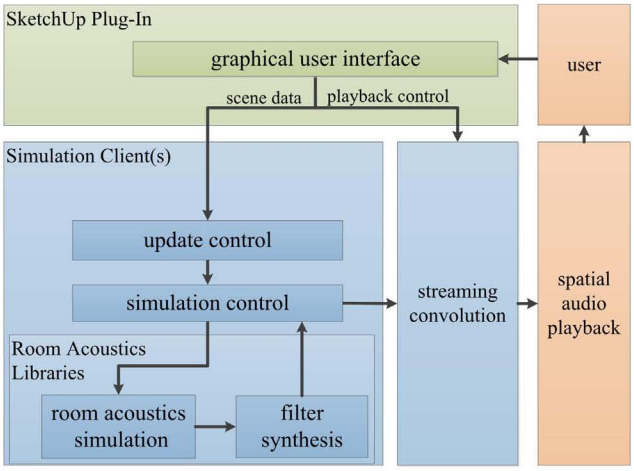


Figure 1: *System components of the auralization system*

3.1. *SketchUp* plug-in

SketchUp[1] is a popular modeling tool, which can easily be learned by beginners and efficiently be operated by expert users. Its functionality can be extended by adding plug-ins using the scripting language *Ruby*. In order to run a room acoustic simulation, the basic functionality of *SketchUp* (setting up a room and assigning wall materials) had to be extended. The implemented plug-in realizes the selection and positioning of sound sources and receivers and automatically sends all acoustically relevant data to the simulation client:

- Room polygons
- Assigned wall materials
- Receiver characteristics
- Source characteristics
- Auralization configuration

A data update is only sent if the situation was changed by the user. Different update rates are used, while source and receiver positions are updated at a high rate (~100 Hz), geometry changes are sent at lower rates (~10 Hz).

To properly modify the situation, the graphical user interface of *SketchUp* had to be extended with special functionality required for the auralization. These extensions are presented in the following subsections.

[1] www.sketchup.com

Sources and Receivers

Besides the room model, for an acoustic transfer path at least one source and one receiver must be defined. Therefore the plug-in introduces additional buttons on the program surface to place acoustic sources and receivers. The objects also contain an adequate visual model (see Fig. 2) and can freely be moved with the *move*-function of *SketchUp*. By right-clicking on the object it is possible to assign a source directivity, an anechoic sound file or soundcard input channel for the playback (sources), and HRTF data (receivers). If multiple receivers are added to the scene, the user has to pick one receiver as the *active* receiver. The sound sources of the scene can be muted and unmuted also by right-clicking on them. To run an auralization, the scene has to contain at least one source. However, adding one *dummy-head* receiver is not essential, because one receiver object is always represented by the position and orientation of the camera viewing the scene. By pressing the button in the toolbar, the user is able to switch between the *camera*- and *dummy-head-receiver* object.

Geometry and Materials

For the modification of the geometry and the wall materials, no extensions were required. Accurate results of the room acoustic simulations can however only be provided if the room does not contain any holes and all materials are assigned on the inside of the room. If the acoustical properties (absorption and scattering) should be considered by the room acoustic simulation, a database file with the same name of the assigned *SketchUp* material has to be included in the material database of the simulation. By pressing a button in the toolbar, the user can visualize the acoustical properties of the selected material (absorption and scattering coefficient for third-octave bands).

Control Panels

The control panel, displayed on the right side in Fig. 2, includes the most important settings of the room acoustic simulation (e. g. Image Source order, Ray Tracing particles) and of the auralization. The simulation components can be switched on and off. Buttons for playback and volume control are included in the toolbar. Another panel, which is currently being developed, enables the user to distribute the workload of the acoustical simulations to multiple simulation clients.

3.2. Simulation Client

Room Acoustic Simulation

The simulation is based on the software library *RAVEN* [16][17], developed at the Institute of Technical Acoustics, RWTH Aachen. The interfaces of the *RAVEN* module include functions to define the scene and run simulations in various configurations. Results of the *RAVEN* simulation models have been validated for various situations, e. g. in [18].

RAVEN defined state-of-the-art algorithms and includes hybrid acoustic simulation models to generate single components of a room impulse response. Fig. 3 shows an energetic room impulse response including the direct sound, the early reflections (calculated by the image sources method) and the reverberation (generated either by a ray tracing algortihm or a statistical artificial reverberation model). The artificial reverberation was integrated

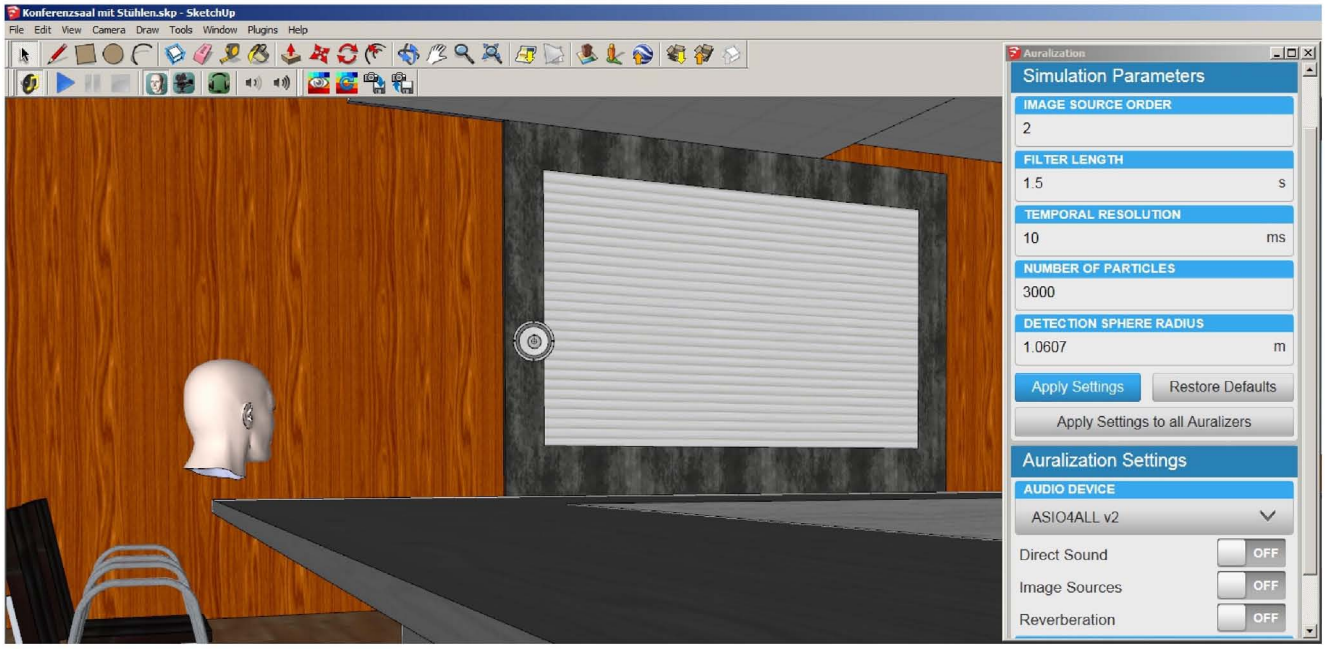


Figure 2: *Graphical user interface of the auralization tool integrated in the modeling software SketchUp*

to demonstrate the difference between and the two reverberation models. As the generation of an artificial reverberation requires significantly less computations in comparison to the ray tracing, it can also be used to provide plausible reverberation feedback without violating the real-time condition in case of only low computational capacities being available. The seperation of the simulation components is reasonable in terms of psychoacoustics, algorithms and data structures [17, 19]. It also offers the possibility to seperately auralize each part of the impulse response.

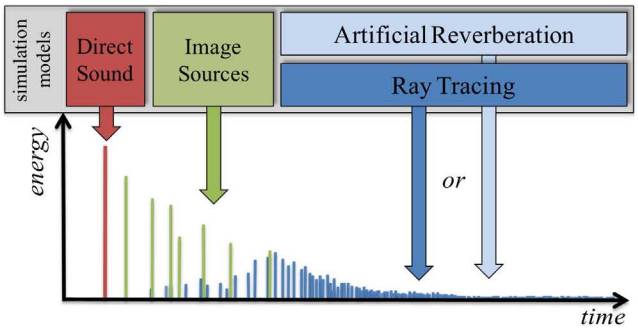


Figure 3: *Simulation models and their contribution to the energetic impulse response*

Simulation Objects

For each active source-receiver combination of the scene, one simulation object is created. For a simulation job, a multi-threading approach is applied, which generates parallel simulation threads according to the configuration of the object (see Table 1). Each simulation component has its own thread, running at different priority levels and update rates. Once one simulation task is finished, the contribution of the simulation result is exchanged in the total room impulse response. Adjusted to the psychoacoustical demands of a highly interative situation (e. g. quick receiver movement in the *camera* mode), the calculation of the direct sound audibility and filter response is executed at a higher priority level than the image sources and the reverberation threads.

Table 1: *Example of a simulation object, generating a binaural room impulse response containing direct sound and reverberation based on a ray tracing*

sourceID	1
receiverID	3
DirectSound	*enabled*
EarlyReflections	*disabled*
ReverberationMode	*Ray Tracing*
FilterMode	*Binaural*

The resulting filter parts of the simulations are cached. If the user enables one simulation component after it has been disabled, it is immediately available. This also applies for switching back and forth between different receiver positions - the room impulse response for the source to the reactivated receiver can be quickly loaded from the working memory without repeating a simulation. This keeps the latency for direct comparisons at a minimum level.

Update handling

A scene modification by the user can affect the auralization in different ways. While a small movement of the source might modify the situation of the direct sound and the early reflections significantly, the perceived reverberation of the room hardly changes.

To classify the modifications made by the user, a class was implemented to analyse the current modification. Fig. 4 shows all possible modification types. *HRTF* and *Directivity* occur if the users selects a different directional characteristic for the receiver or the source in the scene.

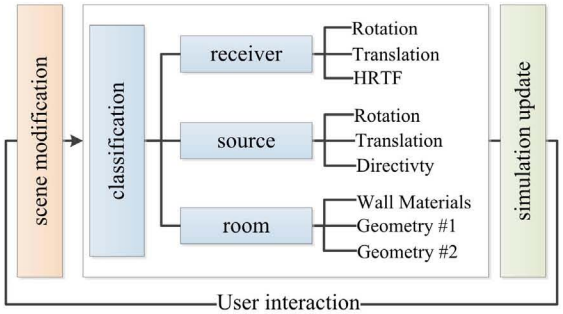


Figure 4: *Classification of a scene modification*

A room modification however represents the most defacing modification. While for an insignificant geometry modification (*Geometry #1*, e. g. dividing a plane of the room without any acoustical effect for the simulation), no update is required and a material exchange only leads to an update of the ray tracing simulation and image source filter, a normal room modification (*Geometry #2*, e. g. shifting a wall or adding objects to the room) necessitates an update of the spatial data structures as well. Spatial subdivision is usually applied to reduce the high number of intersection tests which occur during Geometrical Acoustic simulations [20][21]. For static room geometries with typical polygon counts, the usage of a Binary Space Partioning tree [22] leads to the lowest computation times. For modifiable room geometries however, it has been found that other spatial data structures, e. g. the Spatial Hashing (SH) technique, should be prefered because the update of a spatial hashmap is faster in comparison to the update of the BSP tree[23]. In the here presented auralization tool, both of these spatial data structures are used, according to the situation of the scene. In general, the BSP tree is used to reduce the number of necessary intersection tests within the direct sound, image source and ray tracing calculations. In case of a geometry modification, the calculations of the early parts of the room impulse responses are based on the SH structure, reducing the latency of these calculations significantly.

3.3. Convolution and Streaming

The real-time convolution and streaming module of the auralization tool can be used as a stand-alone application, which can be controlled via a network interface. During the standard application of the tool however, it runs on the same computer as the *SketchUp* software, providing the acoustic feedback through headphones directly to the user. The module supports the convolution and playback of multiple stereo channels, which allows to auralize multiple sound sources at once. The convolution is based on a Fast Fourier Transformation (FFT) convolution algorithm, performing an efficient block-based *Overlap-Save* convolution [24]. Subdividing of the room impulse responses in multiple parts of equivalent length realizes a quick convolution with low latency times. Cross fading is applied to avoid audible artifacts caused by the exchange of filters during the convolution. The headphone playback of the binaural signals is realized using low-latency *ASIO* sound card drivers. Crosstalk cancellation filters for the playback of binaural signals through loudspeakers [25] are currently being integrated.

3.4. Performance Analysis

The performance of the auralization software not only depends on the used hardware, but also on the configuration of the simulation as well as on the complexity of the auralized scene. Instead of an extensive performance analysis for various situations, the calculation times of only one typical example is discussed here. This situation is characterized by the parameters shown in Table 2.

Table 2: *Description of the example scenario*

Room model	*Concertgebouw* Amsterdam
Number of polygons	505
Room volume	20786.3 m^3
Image Source Order	1
Ray Tracing particles	1500 (per octave band)
Radius dection sphere	1 m
Length of Filter	2 s
Reverberation time	2.7 s

The analysed scenario contains one source and receiver, both located inside the concert hall. According to the described parameters and the equations given in [19], the standard deviation of the energy envelope of the late decay is less than $\sigma_L = 0.8\,dB$.

Fig. 5 visualizes the timeline of the most relevant steps in case of a geometry modification. The calculations were carried out on a common desktop computer (*Intel Core i7* @ 3.40 GHz processor, 8 GB RAM).

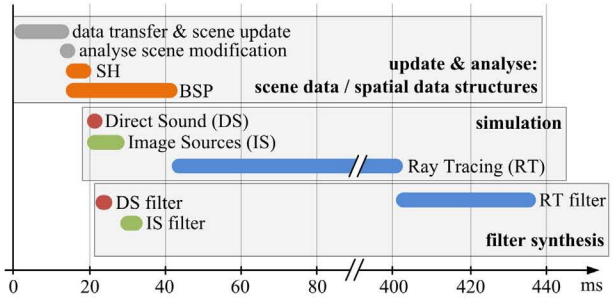


Figure 5: *Timeline of a simulation update after a geometry modification for the given example situation*

As a reaction on the user modifying the geometry, the scene data is updated and the modification is analysed by the simulation client. Both spatial data structures are recalculated in different threads by the geometry controller. The calculation of the Spatial Hashing update takes approximately 3 ms, while it takes 25 ms to update the BSP tree. As soon as the SH update is finished, the simulations of the direct sound and the image sources are executed. Including the binaural filter synthesis, the updated results are sent to the convolution engine after a total processing time of 24 ms (DS) and 34 ms (IS). The simulation thread of the ray tracing algorithm is started as soon as the BSP update has finished. The ray

tracing calculation is *waiting* for the updated BSP-tree because the overall calculation time of a ray tracing based on SH would be significantly higher (700 ms). In a future version, the spatial data structure will be implemented transparently and the ray tracer will already start working on the SH table until the BSP is available. As the binaural filter synthesis of the reverberation is based on an highly accurate approach, which includes spatial information (HRTF data) for each reflection of the room impulse response, the calculation times are relatively high. Different methods for the synthesis of a perceptually equivalent late part of the impulse response are currently investigated [26] and will be implemented. The convolution and reproduction of the binaural signal adds another 5 to 10 ms of delay to the simulation latency, depending on the used buffer size of the sound card driver.

The range of calculation times (one source, one receiver) for similar conditions (configuration, model, hardware) are shown in Table 3.

Table 3: *Range of calculation times for the simulation jobs and the filter synthesis.*

Simulation Component	Simulation Time	Filter Synthesis
Direct Sound	0.5 .. 2 ms	1 .. 5 ms
Image Sources	1 .. 15 ms	1 .. 10 ms
Ray Tracing	250 .. 1500 ms	25 .. 75 ms

A total response time of almost 500 ms violates the real time requirement. However, as the user is not able to perceive minor changes in the reverberation and the most important changes (direct sound and early reflections) can be updated in less than 50 ms even in case of highly interative situations (e. g. flying through the virtual scene in the *camera mode*), the response times are regarded as sufficiently low. Additionally, because typically, a user of the *SketchUp* software is not able of doing more than one geometry modification within one second, a system response including the update of the room's reverberation within one second constitutes *immediate feedback* for the user.

4. APPLICATION

The auralization plug-in is currently successfully being used for demonstrations and exhibitions as well as in university courses including architecture, room acoustics and Acoustic Virtual Reality at different universities. Students are able to design and modify a room while receiving immediate acoustic feedback. This enables them to learn about room acoustics in a playful manner, e. g. exchanging wall materials helps to understand the effects of absorption and scattering. The tool provides a perceptual measure of the room acoustic parameters, which can be calculated and visualized with a tool named *SketchUp-Visualiser* [27], which is also developed at the Institute of Technical Acoustics, RWTH Aachen. A convenient feature of the tool is the investigation of different listening positions by placing multiple receivers at different seats in the audience. By using the *switch-receiver* function, it is possible to quickly switch between different listener positions in a concert hall (see Fig. 6). The dummy heads represent the listening positions, the orange dummy head indicates the currently active receiver.

Figure 6: *Switching receiver positions in a concert hall*

5. CONCLUSION

This work describes the implementation and the possible applications of a real-time auralization tool which is embedded in the popular 3D modeling software *SketchUp*. The development of the graphical user interface was focused on a seamless integration into the software as well as on the intuitive control of the auralization also by non-expert users. The simulation client realizes a state-of-the-art room acoustic simulation, efficiently combining multiple simulation models to create a binaural room impulse responses which is directly processed by the low-latency convolution engine of the tool. The simultaneous use of two spatial data structures is applied to keep the latency of the auralization at a minimum level. Additionally, an elaborated update management reduces the simulation workload in various cases of user interaction.

Although not all simulation results can be calculated in real-time, the auralization tool is able to auralize interactive situations based on physically correct simulation results without significant delays. Because of the parallelized threading concept with a prioritization on quick updates of psychoacoustically dominant parts of the impulse response, the auralization of one sound source in a room can easily be calculated on a single standard desktop computer or laptop, providing an immersive listening experience with multimodal feedback. Currently it is successfully being used for teaching and demonstrations, e. g. in lab courses about room acoustics and spatial hearing. Combined with the calculation and visualization of room acoustic parameters in *SketchUp*, the presented auralization software represents a convenient and reliable tool also for room acoustic consultants.

In future, the usage of multiple simulation clients for the auralization of multiple sources at once will be tested and investigated. Models for priority management in case of larger scenes with multiple sources will be researched and applied to the software. The modular software concept also supports the extension of the software with different simulation and reproduction methods, with Higher-Order Ambisonics and VBAP recently being implemented [28].

6. REFERENCES

[1] J.B. Allen and D.A. Berkley, "Image method for efficiently computing small-room acoustics," *Journal of the Acoustical Society of America*, vol. 65(4), pp. 934–950, 1979.

[2] A. Krokstad, S. Strom, and S. Sörsdal, "Calculating the

acoustical room response by the use of a ray tracing technique," *Journal of Sound and Vibration*, vol. 8, no. 1, pp. 118 – 125, 1968.

[3] Bengt-Inge L. Dalenbäck, "Room acoustic prediction based on a unified treatment of diffuse and specular reflection," *The Journal of the Acoustical Society of America*, vol. 100, no. 2, pp. 899, 1996.

[4] Dirk Schröder, Frank Wefers, Sönke Pelzer, Dominik Stefan Rausch, Michael Vorländer, and Torsten Kuhlen, "Virtual Reality System at RWTH Aachen University," in *Proceedings ICA 2010, 20th International Congress on Acoustics*, 2010.

[5] Nikunj Raghuvanshi, Christian Lauterbach, Anish Chandak, Dinesh Manocha, and Ming C. Lin, "Real-time sound synthesis and propagation for games," *Commun. ACM*, vol. 50, no. 7, pp. 66–73, 2007.

[6] Graham Naylor, "Treatment of early and late reflections in a hybrid computer model for room acoustics," *Journal of the Acoustical Society of America*, vol. 92(4), pp. 2345–2345, 1992.

[7] Mendel Kleiner, Bengt-Inge Dalenbäck, and Peter Svensson, "Auralization-an overview," *Journal of the Audio Engineering Society*, vol. 41, no. 11, pp. 861–875, 1993.

[8] K. Heinrich Kuttruff, "Auralization of impulse responses modeled on the basis of ray-tracing results," *Journal of the Audio Engineering Society*, vol. 41, no. 11, pp. 876–880, 1993.

[9] R. R. Torres, U. P. Svensson, and M. Kleiner, "Computation of edge diffraction for more accurate room acoustics auralization," *J. Acoust. Soc. Am.*, vol. 109, pp. 600–610, 2001.

[10] Michael Vorländer, "International round robin on room acoustical computer simulations," in *International Congress on Acoustics*, 1995.

[11] Ingolf Bork, "A comparison of room simulation software - the 2nd round robin on room acoustical computer simulation," *Acta Acustica united with Acustica*, vol. 86, pp. 943–956, 2000.

[12] Ingolf Bork, "Report on the 3rd round robin on room acoustical computer simulation - Part II: Calculations," *Acta Acustica united with Acustica*, vol. 91, pp. 753–763, 2005.

[13] Jens Blauert, *Spatial Hearing - The Psychophysics of Human Sound Localization*, The MIT Press - Cambridge, Massachusetts, 1996.

[14] B.-I. Dalenbäck and M. Strömberg, "Real time walkthrough auralization - the first year," in *Proceedings IOA Copenhagen*, 2006.

[15] Tobias Lentz, Dirk Schröder, Michael Vorländer, and Ingo Assenmacher, "Virtual Reality system with integrated sound field simulation and reproduction," *EURASIP: Journal on Advances in Signal Processing*, vol. 2007, pp. 187, 2007.

[16] D. Schröder and M. Vorländer, "RAVEN: A real-time framework for the auralization of interactive virtual environments," in *Proc. of EAA Forum Acusticum*, Aalborg, 2011, pp. 1541–1546.

[17] Dirk Schröder, *Physically Based Real-Time Auralization of Interactive Virtual Environments*, Ph.D. thesis, Fakultät für Elektrotechnik und Informationstechnik der Rheinisch-Westfälischen Technischen Hochschule Aachen, 2012.

[18] Sönke Pelzer, Marc Aretz, and Michael Vorländer, "Quality assessment of room acoustic simulation tools by comparing binaural measurements and simulations in an optimized test scenario," *Acta acustica united with Acustica*, vol. 97, no. S1, pp. 102–103, 2011.

[19] Michael Vorländer, *Auralization - Fundamentals of Acoustics, Modelling, Simulation, Algorithms and Acoustic Virtual Reality*, Springer, 2010.

[20] Thomas Funkhouser, Ingrid Carlbom, Gary Elko, Gopal Pingali, Mohan Sondhi, and Jim West, "A beam tracing approach to acoustic modeling for interactive virtual environments," *Proceedings of SIGGRAPH 98*, pp. 21–32, July 1998.

[21] M. Jedrzejewski and K. Marasek, "Computation of room acoustics using programmable video hardware," in *Proc. Computer Vision and Graphics International Conference, ICCVG 2004, Warsaw, Poland*, 2004.

[22] Dirk Schröder and Tobias Lentz, "Real-time processing of image sources using binary space partitioning," *Journal of the Audio Engineering Society*, vol. 54, no. 7/8, pp. 604–619, 2006.

[23] Dirk Schröder, Alexander Ryba, and Michael Vorländer, "Spatial data structures for dynamic acoustic virtual reality," in *ICA2010: 20th International Congress on Acoustics*, 2010.

[24] Frank Wefers and Michael Vorländer, "Optimal filter partitions for real-time fir filtering using uniformly-partitioned fft-based convolution in the frequency-domain," in *Proceedings of the 14th international conference on digital audio effects : September 19-23 : IRCAM, Paris, France / Institut de recherche et coordination acoustique musique (Paris)*. 2011, pp. 155–161, IRCAM-Centre Pompidou.

[25] Tobias Lentz, "Dynamic crosstalk cancellation for binaural synthesis in virtual reality environments," *Journal of the Audio Engineering Society*, vol. 54, no. 4, pp. 283–294, 2006.

[26] Lukas Aspöck, Sönke Pelzer, and Michael Vorländer, "Using spatial information for the synthesis of the diffuse part of a binaural room impulse response," in *DAGA 2014: 40. Deutsche Jahrestagung für Akustik - 10. - 13. März 2014 in Oldenburg*. 2014, pp. 71–72, Deutsche Gesellschaft für Akustik.

[27] Sönke Pelzer, Lukas Aspöck, Michael Vorländer, and Dirk Schröder, "Interactive real-time simulation and auralization for modifiable rooms," *International Symposium on Room Acoustics*, 2013.

[28] Sönke Pelzer and Michael Vorländer, "3d reproduction of room auralizations by combining intensity panning, crosstalk cancellation and ambisonics," in *Proceedings of the EAA Joint Symposium on Auralization and Ambisonics, Berlin.*, 2014.

AURALIZATION AS AN ARCHITECTURAL DESIGN TOOL

Matthew Azevedo[1], *Jonah Sacks*[2]

[1]Consultant in Acoustics,
Acentech Inc.
Cambridge, MA USA
mazevedo@acentech.com

[2] Sr. Consultant in Acoustics,
Acentech Inc.
Cambridge, MA USA
jsacks@acentech.com

ABSTRACT

Auralization provides a valuable tool that allows architects, building owners, and other decision-makers to directly experience the aural implications of design decisions and allows them to make more informed choices. Standard numerical metrics are difficult to relate to aural phenomena without significant practice and frequently fail to capture acoustical issues that are essential to the basic functionality of spaces. Consultants at Acentech have been using auralizations of full soundscapes including many independent sources as design and communication tools for a variety of projects including atria, lecture halls, theaters, and performance spaces. These auralizations have included natural speech and electro-acoustic reinforcement, crowd activity, interactions between PA systems and room acoustics, HVAC noise, wall and window transmission, and the subjective effects of sound masking. In general, clients find the experience of listening to their as-yet unbuilt spaces to be exciting and useful. Though most are not trained listeners, they typically move quickly past the "wow" stage and into critical listening and candid discussion of the different acoustical treatments presented and of the overall sound of the space. This helps architects and project owners to feel connected to the acoustical aspect of the design, and it helps the team to agree on design decisions that may have significant implications regarding cost and aesthetics.

This paper presents several case studies of projects where auralization was an integral part of the design process. Additionally, it describes a rapid auralization design and development process using a MaxMSP-based real-time ambisonic convolution platform.

1 INTRODUCTION

Over the last ten years, consultants at Acentech have worked to develop and refine techniques for using auralization as part of the design process for a variety of buildings and spaces. Much of this work has focused on the auralization of activity sounds in large public spaces such as atrium-type lobby spaces in universities and cultural institutions. The relevant features for auralization of this type of space are:

- overall loudness of activity noise,
- ease of close conversation in the presence of this noise,
- intelligibility of a PA system, and
- transmission of sound among adjoining spaces.

By presenting auralizations of an environment over loudspeakers in an acoustically appropriate meeting room, it is possible for a group of listeners, such as a project design team and owner, to experience the auralization together and to accurately judge many aspects of the auralized space's acoustical character, such as loudness, reverberance, and speech intelligibility.

To create a perceptually realistic soundfield, it is necessary to include multiple source locations and to convolve each source with time-incoherent anechoic sound material. Our early techniques required completing separate convolutions for each source and then combining sources and adjusting their levels in an audio editor. This time-consuming process was replaced by a MATLAB routine that performed all of the convolution and combination while maintaining relative level calibration throughout. But, this process still required that all audio to be presented in the auralization be pre-rendered, resulting in a more time-consuming revision and calibration process and leaving no possibility of level adjustment among sources during playback.

The most recent innovation to this process uses software created with the MaxMSP development environment to perform all of the required convolution in real time during the presentation. This allows for easy toggling on and off of individual sources, level and timbre manipulation of individual sources, and switching among various architectural design conditions, all during the presentation.

This paper presents case studies of several auralizations that employed these techniques in different ways to demonstrate relevant information about projects and to provide direct experience of the acoustics of project spaces.

2 A REAL-TIME CONVOLUTION PLATFORM FOR AURALIZATION DEVELOPMENT

We conceptualize auralization development as four main processes:

- modeling,
- impulse generation,
- source material selection and convolution, and
- playback calibration.

Typically, our modeling begins in SketchUp and is then exported to CATT Acoustic where material properties are added. The CATT model is then processed with TUCT, CATT's room analysis subprogram which generates impulse responses.

Because CATT did not provide convolution and summing of different anechoic audio material with multiple sources, we developed a MATLAB script in 2008 which convolved source

material with first-order ambisonic (B-Format) impulses responses for each sound source in the model, and then summed and rendered the resulting audio as either a four channel (quadraphonic) or five channel (5.1 surround) WAV file using simple decoding filters generated by CATT Multivolver. This process was effective, but revisions and calibration were time-consuming since any changes involving source material required repeating the entire convolution and summing process. This placed auralization out of reach for most project budgets.

2.1 Auralization with pre-rendered audio: Museum Atrium

The first major step in the early development of our auralization program took place in 2003 through 2005 during design of a major new wing at the Museum of Fine Arts Boston, which opened to the public in 2010. This work was presented previously at the EAA Symposium on Auralization in Espoo, Finland in 2009 [1] and at Internoise 2009 in Ottawa, Canada in 2009 [2]. The auralization presented a familiar architectural acoustics problem and solution: excessive loudness in a large, reverberant public space, addressed with the inclusion of large areas of acoustically absorptive material. The design team and museum trustees heard a predictive auralization of their future space during a fundraising banquet, complete with 500 talkative diners, a live swing band, and an amplified speech by the museum's director. Participants experienced the difficulty of speaking with their table-mates at this busy function in the absence of acoustical treatment and registered relief when the recommended treatment was added in the virtual environment. The auralization included two source locations for diners, a third for the band, and a fourth for public address loudspeakers. Three architectural design conditions were presented: no acoustical treatment, a small amount of acoustical treatment, and the recommended amount of acoustical treatment. All audio was pre-rendered and played back as four-channel wav files. Anechoic sounds were gathered from various sources, including original recordings of conversation and clearing of dishes made in our nearly-anechoic presentation room, speech from museum audio tour recordings, and studio recordings of a swing band.

Figure 1: The Ruth and Carl J. Shapiro Family Courtyard at the Museum of Fine Arts Boston, MA USA.

The Ruth and Carl J. Shapiro Family Courtyard opened in 2010, and the response from the museum and the public has been overwhelmingly positive. The courtyard performs acoustically as designed and supports a wide range of uses. Subjectively, it sounds remarkably similar to the auralization: live but well-controlled even when full of activity.

This auralization, though successful, was time-consuming to produce. Minor adjustments to relative levels required repeating much of the process. Such adjustments are often necessary in this kind of auralization, as different source material is recorded and calibrated at different levels. Final relative level balancing is often an iterative process, checked with the aid of a sound level meter, CATT output data, and various reference sources. When balancing, for example, a live band with activity noise from diners, we assume a role similar to that of a live sound board operator, adjusting the band to a subjectively appropriate level. The difficulty of making such adjustments was a source of frustration to us for several years.

2.2 Towards a more flexible and efficient auralization development platform

Two points drove the redevelopment of our earlier auralization process. First was a desire to reduce the time and cost required to create an auralization presentation, so that we could make wider use of auralizations in our consulting practice. Second was a need to maintain design flexibility for as long as possible in the development process, so as to respond efficiently to client requests for changes.

The first two steps of our auralization process, modelling and impulse generation, are by far the most time intensive. For complicated auralizations with many sources, processing the model can require several CPU-weeks of processing time. While multicore CPUs and smart model design which accommodates parallel processing significantly reduces total processing time, processing a complicated model still frequently requires tying up multiple computers for several days. Because of this, we desired a platform where the model only needs to be processed once, and most needed tweaks or adjustments can happen downstream.

Convolution in MATLAB is a fairly quick process. However, our process of pre-rendering the auralization meant that any change, no matter how small, required a full repetition of the convolution and summing steps – not a trivial matter when making final preparations on the day of a client meeting!

These issues pointed towards a platform that could perform the convolution and ambisonic decoding in real-time, while also providing flexible options for additional processing.

2.3 Real-time multichannel convolution using MaxMSP

We currently produce all of our auralizations using the MaxMSP programming environment published by Cycling ’74 for all steps following the generation of impulse responses in TUCT. These include convolution, ambisonic manipulation, and the final presentation. Max was originally created as a graphical, object-oriented programming platform for musical composition, and has been steadily upgraded to include a robust set of audio (MSP, Max Signal Processing) and video (Jitter) processing tools. Max also makes graphic user interface (GUI) implementation trivial, since many of the graphical function objects are directly useable as UI objects as well. One of Max’s strengths is that it is an open

platform that allows independent development of additional function objects (called "externals"); this extensibility is critical when using Max as an auralization tool since many of required functions, such as convolution and ambisonic encoding and decoding, are not part of the base MaxMSP object library.

The most important tools for auralization in Max are the convolution externals in the HISSTools library, written by Alex Harker and Pierre Alexandre Tremblay of the Huddersfield Immersive Sound System (HISS) research group at the University of Huddersfield, UK. The HISSTools Impulse Response Toolbox provides simple yet powerful objects for real-time convolution in Max, as well as tools for measuring and manipulating impulse responses. HISSTools can be freely downloaded from the HISS website [3].

We use two libraries for ambisonic processing in our auralizations, *Ambisonics Externals for MaxMSP* written by Philippe Kocher and Jan Schacher at the Institute for Computer Music and Sound Technology at the Zurich University of the Arts, Zurich, CH and *Ambisonic tools for Max/MSP* written at the Center for New Music & Audio Technologies at the University of California at Berkeley, Berkeley, CA, USA. Both of these libraries are freely available at the websites of the respective organizations [4, 5]. These tools allow decoding the convolved ambisonic signals for playback on arbitrary speaker arrangements and also allow sound sources to be placed and moved within an ambisonic soundfield.

Using MaxMSP as a development platform has dramatically reduced the time required to create a complex auralization and allows us to retain the flexibility to modify the auralization all the way up to, and in some cases during, the final presentation to our clients. This in turn leads to a more responsive and deeper engagement with our clients resulting in better communication between all parties and more successful projects.

3 CASE STUDIES IN AURALIZATION-DRIVEN DESIGN

3.1 Multiple concurrent sources: Concert Hall

The new General Academic Building at the University of Massachusetts Boston, currently in design, will include a 400-seat music recital hall to be used for solo recitals, chamber music, jazz combos, orchestra, vocal chorus, and other musical ensembles. The design includes acoustical variability in the form of curtains at the lower walls. An auralization presented several different types of performance, including solo piano, jazz band, symphony orchestra, and vocal chorus, heard at three audience positions, with various curtain configurations. For the large ensembles, two source locations were included on stage to provide some left-right spatial spread. Stereo anechoic material was used where available, and in other cases monaural material was used. The source material was taken from the Denon compact disc "Anechoic Orchestral Music Recording" (1995) and from Wenger Corporation's "Anechoic Choral Recordings" (2004). The auralization helped the university music faculty to gain confidence in the design, particularly its acoustical variability.

Figure 2: CATT Acoustic model of the University of Massachusetts recital hall, Boston, MA, USA

We have begun experimenting with using anechoic recordings of individual instruments made by researchers at Aalto University [6], and hope to use these to render orchestral music with improved spatial accuracy in future projects.

3.2 Natural source, sound reinforcement, and HVAC system noise: Multifunction Hall

Margery Milne Battin Hall at the Cary Memorial Building in Lexington, Mass. is a multifunction auditorium built in 1928. The hall's users were unsatisfied with its sound reinforcement system, and the building's HVAC system was loud and inefficient. Acentech was asked to design an upgraded PA system and to assist the mechanical engineer in quieting the ventilation system.

To demonstrate the predicted effects of our various design recommendations, and to help the client prioritize their use of limited funds, we created and presented an auralization. This auralization included a natural sound source on the stage, the HVAC background noise, the full existing sound system (12 sources), the full proposed sound system (21 sources), and room responses with and without acoustically absorptive curtains. Several options were included for HVAC noise. The current noise was presented based on the system noise level and spectrum measured in the auditorium, and two proposed upgrades to the HVAC system were presented based on levels and spectra calculated according to ASHRAE guidelines. Since the auralization is being processed in real time, the background noise, PA system, and curtains can be toggled during playback with a continuously running source. This allowed for seamless comparisons between the different options and made the differences much easier for the clients to understand.

Figure 3: Battin Hall, Cary Memorial Building, Lexington, MA USA

Because the convolution takes place with first order ambisonic impulse responses, each source requires four channels of real-time convolution. While we have successfully had over 50 channels of convolution running concurrently on a quad-core CPU, the 140 channels of convolution that would be required to generate the responses of all 35 sources in real time is not practical. To reduce the number of concurrent convolution channels, we grouped all of the loudspeaker sources for the existing PA system and the sources for the proposed PA system, which will be fed the same anechoic source material, into a single TUCT run and then exported a summed impulse response of the entire group. This allowed us to present all 35 sound sources while only requiring 16 concurrent channels of convolution.

3.3 Custom source material and integration of measured impulse responses: Aquarium Exhibit Hall

One of the New England Aquarium's most popular exhibits is the Shark and Ray Touch Tank, where guests can pet small sharks and stingrays. The space has a naturally high background noise level due to the tank's pumps and the electric dryers for guests to use after washing their hands. When combined with the added noise of as many as 100 excited elementary school students, the hall containing the Touch Tank can be extremely loud. The New England Aquarium was interested in ascertaining the noise impact on the animals in the tank, as well as exploring potential room treatments to reduce the overall noise level for the comfort of their staff and guests.

Figure 4: The Touch Tank at the New England Aquarium, Boston MA

For the auralization of this space to be successful, it was critical that appropriate audio source material was used. Since the audience for this auralization was intimately familiar with the modelled environment the sounds presented had to match their expectations in order to be convincing. To achieve this, we recorded many sounds within the space, including the background sound of the unoccupied space, the hand dryers, and the sound of excited children. The speech of one of the educators who directs patrons in safely touching the animals was transcribed and re-recorded anechoically. Since we were also concerned about the noise impact on the animals in the water, we also recorded the background sound underwater using a hydrophone, and measured the transfer function impulse response between a microphone just above the water and a hydrophone in the tank.

The auralization was very successful, but in an unusual way: in this case we demonstrated that additional acoustical treatment would not make a substantial difference in the loudness of the space, and that the most effective option for quieting the exhibit would be to control the number of patrons in the space at one time. We also clearly demonstrated that the airborne noise generated by patrons in the exhibit was almost entirely masked inside the tank by the noise of the pumps that circulate the water. By making a modest investment in the auralization, the Aquarium was able to avoid making a large investment in room treatments that would have required shutting the exhibit down to install and then been ineffective in addressing their noise concerns.

This auralization served to highlight a common thread in all of our auralization work: the importance of appropriate source material in creating perceptual veracity (as opposed to parametric accuracy) in an auralization. By having access to a semi-anechoic space, we are able to record custom material for use in our auralization work to ensure that the character of the source material is appropriate to the expectations of our audience. By tailoring the source material to match what our clients hear in their current spaces, listening to the auralization can be focused on the parametric aspects of the presentation without being sidetracked by clients' being disengaged from the listening experience by the distraction caused by inappropriate audio content.

3.4 Complex soundscape with varied acoustics in a large space: University Atrium

The centrepiece of the new home of the Olin Business School at Washington University in St. Louis, MO is the Forum, an amphitheatre-like lecture and presentation space which extends through several floors of circulation space and is ultimately open above to a five storey glass, wood, and stone atrium featuring a café with seating all around the opening. Despite our presentation of calculated reverberation times and expected background sound and speech intelligibility levels, the client remained unsure that the extensive acoustical treatments we recommended were truly needed. After further discussion, it was decided that an auralization would be the best way for the client to make an informed decision regarding room treatments, and to develop appropriate expectations for the acoustical performance of this complex space. Of particular interest to the client was the level of café activity noise that would be audible in the Forum during a lecture presentation.

This auralization included a speech source and sound reinforcement system in the Forum, and four independent activity sources in the atrium. Many different types of anechoic source material were used for the activity sources, ranging from quiet studying with light footfall noise to boisterous conversation and the sounds of eating and tables being cleared. The level of ambient activity can be varied in real time, and for each activity level chosen by the operator appropriate activity samples are dynamically selected and fed into the convolution engines for the various modelled activity sources. A close-mic'd recording of a speech by an Olin School professor was used as the sound source in the Forum, which was both context-appropriate and helped to anchor the audience's sense of place in the modelled environment. This recording was made by the school for public relations

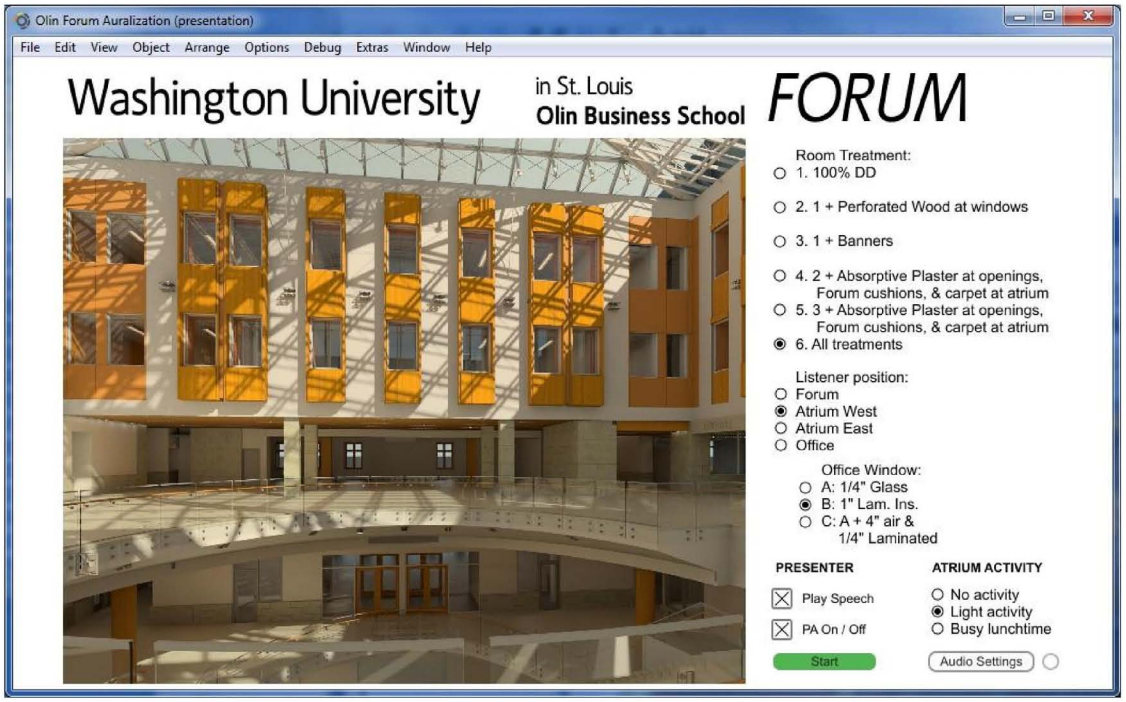


Figure 5: The GUI (Graphical User Interface) of the Olin Business School Forum auralization, constructed in MaxMSP.

purposes and, while not anechoic, was sufficiently dry to be effective for this auralization of a reverberant space. Moreover, the cardioid podium microphone used for the original recording is of the type which speakers will use in the built space, which, while compromising the timbre of the unamplified speech, better presents the full signal path of amplified speech.

Five varying levels of acoustical treatments were modelled, ranging from the initial design to fully treating every available surface. Five listener locations are available, including one inside a faculty office overlooking the atrium which incorporates the transmission loss of the window glazing and allows for three different glazing types to be auditioned. Of course, all of these parameters are freely variable at run time without a need to stop and restart the source recording. See Figure 5 for an illustration of the graphical user interface (GUI) used to select the various options available in this auralization.

In this case, school administrators in attendance were able to make the connection between the architectural design, the numerical descriptions of the room acoustics, and the percept of experiencing the described environment through the auralization in a way that they had not been able to from a written report. Having come into the auralization sceptical about the need for the extent of acoustical treatments that we'd recommended, their final decision was that fully treating the Forum and atrium was a requirement for the success of the project. They were also alerted to practical considerations regarding scheduling of Forum presentations and the need to control Café activity during these presentations.

3.5 Algorithmically generated, massively multichannel outdoor soundscape: 17th Century Churchyard

While not an architectural design problem *per se*, The Virtual Paul's Cross Project [7] represents the current cutting-edge of Acentech's auralization program. This project was a joint investigation led by Dr. John Wall at North Carolina State University which included the English, architecture, and linguistics departments at NCSU, the St. Paul's Cathedral archaeology staff, and acoustical consultants at Acentech. The goal of the project was to recreate the soundscape and visual surroundings of John Donne's 1622 Gunpowder Day sermon at Paul's Cross in the churchyard outside of St. Paul's Cathedral as it was prior to the Great Fire of London in 1666.

The Paul's Cross auralization includes the modelled acoustics of the open churchyard and its surrounding buildings, a 2 hour and 15 minute anechoically-recorded sermon, and an "artificially intelligent" crowd which is variable in real time from between zero to 5000 people and that "listens" to the sermon and then selects appropriate reactions from a custom-produced sample library. Birds fly overhead, horses trot past at the edges of the crowd, dogs bark, and the bells of St. Paul's mark the quarter hours. All of this is observable from twelve different listener locations, and crowd size and listener location are variable during playback without any interruption to the sermon.

The ability to generate the full soundscape in real time was absolutely critical to the success of this project. If we had been required to hand-arrange all of the crowd reactions and ambient sound events, we would have been faced with a task comparable to doing sound design for a feature-length film which would have required an amount of work drastically outside the project budget. By using statistical models to drive the auralization, we

were able to specify model parameters and then automatically generate all of the ambient sounds, with that added benefit that no two listens through the sermon are exactly the same. Also, not needing to pre-render the audio substantially reduced the size of the data assets. As an example, the audio assets for the auralization total 2.6 GB of data. If all configurations of the auralization were to be pre-rendered, it would result in 240 GB of data, almost a hundredfold increase. As we have seen in other cases, real-time auralization allows the sermon to run continuously while model parameters are changed, resulting in a notably more enveloping and believable aural experience for the listener.

While triggering the more static environmental sounds of wind, animals, and church bells was fairly straightforward, generating a crowd that could track the dynamics of the sermon was a greater challenge. A rudimentary artificial intelligence (AI) was written in Max that listens to the sermon and then selects a sample from the library of crowd recordings of an appropriately intense response. The full crowd is made up of ninety independent instances of this AI, which are tiled across the soundfield of the churchyard. Each AI's behaviour preferences are randomized at runtime. This allows the impression of a very large crowd of independent listeners to be generated from a small amount of source material.

To fully model each node of the AI and environmental sound within CATT would have required 250 days of processing time and 400 additional channels of real-time convolution. Instead, a simplified approach was taken. The dry output of each AI and environmental sound generator was encoded into an ambisonic signal to give the direct sound the appropriate spatial and level cues in the final presentation. Then, the omnidirectional channels (W in B-Format parlance) of all of the AIs and effects were summed together and convolved with a set of impulses derived from an omnidirectional source in the center of the churchyard with the direct sound removed. Thus, the direct sound is presented in a spatially accurate way via the ambisonic encoding of the dry audience and effects, and the reverberant sound is still perceived as coming from reflective surfaces that are accurately placed around the listener. Since precise spatial localization was desired for the preacher and the church bells, they are auralized with dedicated sources with appropriate directional characteristics in the model. Even with the shortcuts for reducing the channel load of the ambient sound sources, the Paul's Cross auralization requires over 100 concurrent audio playback channels (but only 12 convolution channels, due to the simplifications described above) when the crowd size is set to its maximum.

The Virtual Paul's Cross auralization has facilitated a new level of conversation and inquiry into Donne's preaching and more broadly of the experience of the churchgoing public of 17th century England. We have been able to provide a direct experience of the way an unamplified voice interacts with an outdoor forum such as the historical Paul's Yard and of how speech intelligibility and loudness change for many listener positions and crowd sizes. Being able to switch between listener positions and crowd sizes immediately and in real time allows for much easier and more natural comparisons to be drawn from the auralization than would be possible from tabular data and opens the experience up to a wide range of interested listeners who lack the acoustical knowledge to gain a meaningful understanding of aural events from technical descriptions of them.

Figure 6: A rendering of Paul's Cross, ouside of St. Paul's Cathedral, London, UK

4 CONCLUSIONS

We have found auralization to be uniquely capable of bridging the "communication gap" between acousticians, architects, and project stakeholders in our consulting practice. Allowing parties to come together and listen as a group to the acoustical implications and possibilities of architectural design decisions facilitates understanding in a powerful way that allows for rapid and harmonious decision making. In many cases, auralization is not only the best way to communicate acoustical information, but can result in overall cost savings to a project by reducing the need for revisions during the design process or renovations to correct problems after construction.

In particular, a real-time auralization system allows for richer and more immersive auralizations, without interrupting the flow of the source material and an immediate response of the sound to interactions with the presentation GUI. A real-time system also results in substantial time and cost savings to the client, and allows both initial development as well as revisions as the project progresses to happen on a schedule consistent with the aggressive timelines of many architectural projects.

5 REFERENCES

[1] I. Pieleanu, "Auralizations of Public Atrium Spaces", *Proc. Of the EAA Symposium on Auralization, Espoo, Finland, 15-17 June 2009*, 2009.

[2] J. Sacks, "Auralization for Public Spaces," *Proc. Internoise 2009, Ottawa, Ontario, Canada, August 23-26 2009,* 2009.

[3] HISS: Huddersfield Immersive Sound System, (University of Huddersfield), http://www.thehiss.org/ , accessed 25 Nov. 2013.

[4] Ambisonics Externals for MaxMSP, (Institute for Computer Music and Sound Technology, Zurich University of the Arts), http://www.icst.net/research/downloads/ambisonics-externals-for-maxmsp/ , accessed 15 Jan. 2013.

[5] Ambisonic tools for Max/MSP, (Institute for Computer Music and Audio Technologies, University of California at Berkeley), http://cnmat.berkeley.edu/link/4415 , accessed 15. Jan. 2013.

[6] T. Lokki, J. Pätynen, “Applying Anechoic Recordings in Auralization”, *Proc. Of the EAA Symposium on Auralization, Espoo, Finland, 15-17 June 2009*, 2009.

[7] B. Markham, M. Azevedo, J. Wall, "Acoustical archaeology - Recreating the soundscape of John Donne's 1622 gunpowder plot sermon at Paul's Cross." *The Journal of the Acoustical Society of America* 133.5 (2013): 3581-3581.

EVALUATING THE ACCURACY OF THE AMBISONIC REPRODUCTION OF MEASURED SOUNDFIELDS

Sam Clapp
Graduate Program in Architectural Acoustics
Rensselaer Polytechnic Institute
Troy, New York, USA
clapps@rpi.edu

Anne Guthrie
Arup Acoustics
New York, NY
anne.guthrie@arup.com

Jonas Braasch
Graduate Program in Architectural Acoustics
Rensselaer Polytechnic Institute
Troy, New York, USA
braasj@rpi.edu

Ning Xiang
Graduate Program in Architectural Acoustics,
Rensselaer Polytechnic Institute
Troy, New York, USA
xiangn@rpi.edu

ABSTRACT

A spherical microphone array can encode a measured soundfield into its spherical harmonic components. Such an array will be subject to limitations on the highest spherical harmonic order it can encode and encoding accuracy at different frequencies. Ambisonics is a system designed to reproduce the spherical harmonic components of a measured or virtual soundfield using multiple loudspeakers. In ambisonic systems, the size of the sweet spot is wavelength dependent, and thus decreases in size with an increase in frequency. This paper examines how to reconcile the limitations of the recording and playback stages to arrive at the optimum ambisonic decoding scheme for a given spherical array design. In addition, binaural models are used to evaluate these systems perceptually.

1. INTRODUCTION

Spherical microphone arrays have been studied extensively for beamforming [1, 2], source localization, and other applications. Likewise, much theoretical and practical work has been done on the optimum methods for ambisonic decoding [3, 4, 5]. Both systems use the concept of spherical harmonics: spherical microphone arrays can decompose a soundfield into its spherical harmonic components, while ambisonic decoding can reconstruct the spherical harmonic components of a soundfield for a listener.

The performance of spherical microphone arrays is frequency-dependent, affected primarily by the array's size and number of sensors. In many applications, such as beamforming and direction-of-arrival (DOA) estimation, a narrow frequency band is used, where the array's performance is optimum. However, restricting oneself to a narrow frequency band is untenable when presenting auditory scenes to listeners. Thus, we require a way to utilize information from outside of the optimum band, where higher spherical harmonic orders might not be accurately decomposed, but lower orders are.

Much of the literature on ambisonics deals with the reproduction of simulated soundfields, where the exact DOA of every sound event is known, and the restrictions on the highest spherical harmonic order come only from the number of channels in the loudspeaker array.

This paper examines a method for determining mixed-order ambisonic decoding schemes determined by the constraints of the two systems, as well as a way to evaluate the perceptual accuracy of these schemes using binaural models.

2. SPHERICAL HARMONICS

The homogeneous wave equation is given in its general form by:

$$\nabla^2 p = \frac{1}{c^2}\frac{\partial^2 p}{\partial t^2}. \tag{1}$$

If expressed in spherical coordinates, solutions can be obtained through a separation of variables, yielding sets of functions for the radial, azimuthal, and elevation components [6]. Solutions to the azimuthal component are given by either sine and cosine terms or complex exponentials, while solutions to the elevation component are given by the associated Legendre functions ($P_n^m(x)$) in the cosine of the elevation angle. Combining these two components (and adding a normalization term) yields the (complex-valued) expression for spherical harmonics:

$$Y_n^m(\theta,\phi) = \sqrt{\frac{2n+1}{4\pi}\frac{(n-m)!}{(n+m)!}} P_n^m(\cos\theta) e^{im\phi}. \tag{2}$$

(The real-valued expression uses sine and cosine terms for the azimuthal component.)

The spherical harmonics form an orthonormal basis on the sphere:

$$\int_0^{2\pi}\int_0^{\pi} Y_{n'}^{m'}(\theta,\phi)^* Y_n^m(\theta,\phi)\ \sin\theta\, d\theta\, d\phi = \delta_{nn'}\delta_{mm'}. \tag{3}$$

3. SPHERICAL MICROPHONE ARRAY PROCESSING

3.1. Spherical Harmonic Decomposition

When a plane wave impinges upon a rigid sphere, the sphere will radiate a spherical wave whose intensity will vary as a function

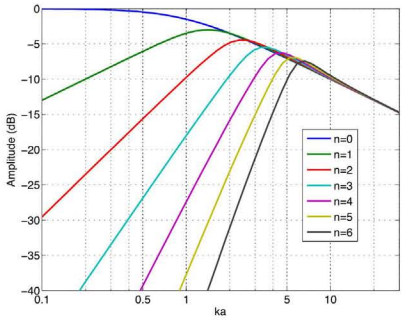


Figure 1: Modal amplitude for orders zero through six, as a function of ka, a quantity that relates the wavelength of the incoming plane wave to the radius of the spherical array.

of the incident wave's angle of incidence and wavelength (in relation to the radius of the sphere). It is shown in [6] and [7] that by solving the wave equation with the appropriate boundary conditions, we arrive at an expression for the pressure at a point on a rigid sphere of radius a (denoted by its angular position (θ, ϕ)) due to a plane wave incident from (θ_i, ϕ_i) with amplitude P_0 and wavenumber $k = 2\pi f/c$:

$$p(\theta, \phi, ka) = 4\pi P_0 \sum_{n=0}^{\infty} i^n b_n(ka) \sum_{m=-n}^{n} Y_n^m(\theta, \phi) Y_n^m(\theta_i, \phi_i)^*, \tag{4}$$

with

$$b_n(ka) = j_n(ka) - \frac{j_n'(ka)}{h_n^{(1)'}(ka)} h_n^{(1)}(ka), \tag{5}$$

a quantity referred to as the modal amplitude, which is shown for several orders in Fig. 1. The slope of each curve is 3 dB per octave multiplied by the order.

Now, if we apply a weighting factor W to each point on the sphere:

$$W_{n'}^{m'}(\theta, \phi, ka) = \frac{Y_{n'}^{m'}(\theta, \phi)^*}{4\pi i^{n'} b_{n'}(ka)}, \tag{6}$$

and then integrate over the entire sphere, we can use the orthonormality of the spherical harmonics (Eq. 3) to yield the following result:

$$\int_0^{2\pi} \int_0^{\pi} W_{n'}^{m'}(\theta, \phi, ka)\, p(\theta, \phi, ka)\, \sin\theta \, d\theta \, d\phi = P_0 Y_{n'}^{m'}(\theta_i, \phi_i)^*. \tag{7}$$

Thus, we can determine the spherical harmonic components of the incident plane wave. However, evaluating the integral in Eq. 7 requires a continuous spherical transducer, and in practice, we must sample the pressure at Q discrete points on the sphere (denoted by their angular positions (θ_q, ϕ_q)), leading to the following summation:

$$\sum_{q=1}^{Q} W_{n'}^{m'}(\theta_q, \phi_q, ka)\, p(\theta_q, \phi_q, ka)\, C_n^m(\theta_q, \phi_q) \approx P_0 Y_{n'}^{m'}(\theta_i, \phi_i)^*, \tag{8}$$

where C_n^m are the quadrature coefficients. Using a nearly uniform sampling scheme, we can estimate the spherical harmonic components up to order N such that $Q \geq (N+1)^2$.

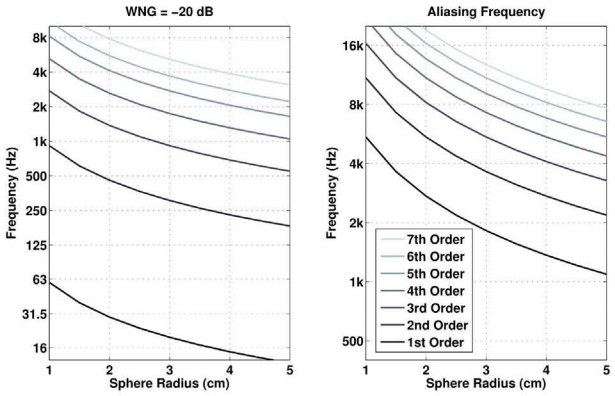


Figure 2: Left side: frequency at which WNG = -20 dB as a function of sphere radius for a 64-channel array at different spherical harmonic orders. Right side: frequency at which aliasing errors occur as a function of sphere radius for different spherical harmonic orders.

3.2. Error Sensitivity

Spherical microphone arrays lend themselves well to beamforming, a set of techniques for multi-channel sensors that allow for spatial filtering [2, 8, 9, 10]. These techniques involve solving for a set of weighting factors that are applied to each channel on the array, allowing the array to "look" in a particular direction at a particular frequency. The robustness of the beamformer is given by a quantity known as the white noise gain (WNG):

$$\mathrm{WNG}(\theta_0, \phi_0, \theta_q, \phi_q, ka) = 10 \log_{10} \left(\frac{|\mathbf{d}^T \mathbf{W}|^2}{\mathbf{W}^H \mathbf{W}} \right), \tag{9}$$

where $\mathbf{d}^T$ is a column vector of the microphone pressure values due to a plane wave impinging on the sphere from some direction (θ_i, ϕ_i), and $\mathbf{W}$ the sensor weights to look in that direction. WNG represents the array's sensitivity to noise and microphone positioning errors, with negative values representing an amplification and positive values representing an attenuation of (spatially uncorrelated) white noise. A spherical array is most sensitive to these types of errors at lower frequencies, where higher order spherical harmonic components are low in level and must be amplified considerably. The lefthand portion of Fig. 2 shows the frequencies at which WNG = -20 dB for a 64-channel array of varying radius. This value can be used as a guideline to determine the lowest frequency at which a certain order of spherical harmonic components can be accurately measured by a given array.

3.3. Aliasing

At higher frequencies, the aliasing of higher order spherical harmonic components into lower orders becomes an issue. We can see from Eq. 4 that a plane wave is not spherical harmonic order-limited and from Eq. 5 that higher-order components become more prominent at higher frequencies. Thus, aliasing errors affect the array for frequencies such that $ka > N$, where a is the radius of the sphere and N is the highest spherical harmonic order that can be measured by the array [11]. This frequency is shown in the right-hand portion of Fig. 2 for a range of sphere radii and spherical harmonic orders, with arrays of smaller radii and higher order (i.e.

with a greater number of channels) offering higher thresholds for aliasing.

4. AMBISONICS

4.1. Basic Decoding

Ambisonics is a system that uses multiple loudspeakers to synthesize a soundfield, where the gains of the loudspeakers are determined based on the spherical harmonic expansion of the soundfield [12, 3, 4, 5].

Let us start with a plane wave of wavenumber $\mathbf{k_i}$ and incident from the direction (θ_i, ϕ_i), defined by its spherical harmonic coefficients as:

$$p = e^{i\mathbf{k_i r}} = 4\pi \sum_{n=0}^{\infty} i^n j_n(kr) \sum_{m=-n}^{n} Y_n^m(\theta,\phi) Y_n^m(\theta_i,\phi_i)^*. \tag{10}$$

If we want to synthesize this plane wave with L plane wave sources located at (θ_l, ϕ_l) for $1 \leq l \leq L$, each with an amplitude of w_l, then our expression for the synthesized field $\hat{p}$ is:

$$\hat{p} = 4\pi \sum_{n=0}^{\infty} i^n j_n(kr) \sum_{m=-n}^{n} Y_n^m(\theta,\phi) \sum_{l=1}^{L} w_l Y_n^m(\theta_l,\phi_l)^*. \tag{11}$$

Then we solve for each n and m in order to have Eq. 11 be equal to Eq. 10:

$$\sum_{l=1}^{L} w_l Y_n^m(\theta_l,\phi_l)^* = Y_n^m(\theta_i,\phi_i)^*. \tag{12}$$

These amplitudes w_l can be solved for up to order N, subject to the requirement that $L \geq (N+1)^2$ for 3-D arrays and $L \geq 2N+1$ for horizontal arrays. This is known as basic decoding.

4.2. Wavefield Error Analysis

One method of evaluating the quality of the reconstruction is by calculating the normalized radial error between the synthesized sound field $\hat{p}$ and the sound field being reconstructed, p, given by:

$$\bar{\epsilon}(kr) = \frac{\int_0^{2\pi}\int_0^{\pi} |p-\hat{p}|^2 \sin\theta \, d\theta \, d\phi}{\int_0^{2\pi}\int_0^{\pi} |p|^2 \sin\theta \, d\theta \, d\phi}. \tag{13}$$

This quantity is a function of the distance from the center of the array as related to the wavelength one is examining, expressed as kr, and is shown for 1st through 7th order (moving from left to right) in the top portion of Fig. 3. A given value of kr will be a larger distance from the center of the array for a lower frequency (i.e. longer wavelength) than for a higher frequency. Thus, the sweet spot will be larger for lower frequencies than for higher frequencies.

One way to evaluate the performance of ambisonic systems in terms of human perception is to examine the frequency at which the radius of the sweet spot (as determined by some threshold in the normalized radial error) becomes smaller than the radius of the average human head (given in [4] as 8.9 cm). As the order of ambisonic reproduction increases, so will this frequency, as shown in the bottom portion of Fig. 3, for several different thresholds of normalized radial error. Thus, as higher spherical components are added to the reconstruction, higher frequencies are reconstructed accurately at the two ears.

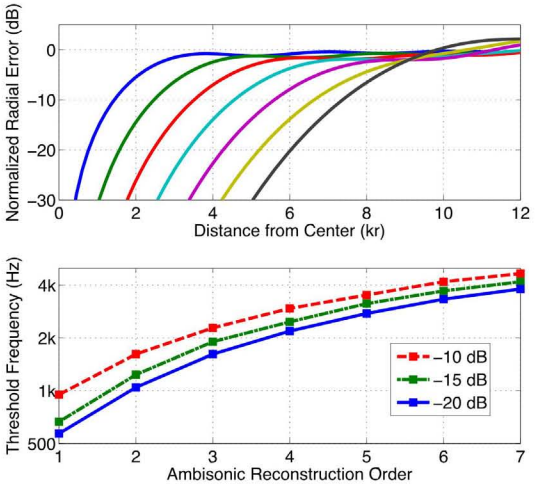


Figure 3: The top shows normalized radial error (in dB) as a function of distance from the center of the array (in units of kr), from 1st to 7th order basic decoding (moving left to right), after Ward and Abhayapala[13] and Poletti[4]. The bottom shows the frequencies at which the -10, -15, and -20 dB error thresholds occur at a radius of 8.9 cm, the size of a typical human head.

4.3. Max-r_E Decoding

For a given order of ambisonic reproduction, there will be some frequency above which the area of the sweet spot will be smaller than the size of a typical human head. Thus, audio content above this frequency will not be simulated accurately at the listener's ears. In [14], Gerzon proposed that high frequency localization can be predicted by the direction of the energy vector at the center of the array, $\hat{\mathbf{r}}_\mathbf{E}$, as given in:

$$r_E \hat{\mathbf{r}}_\mathbf{E} = \frac{\sum_{l=1}^{L} w_l^2 \hat{\mathbf{u}}_\mathbf{l}}{\sum_{l=1}^{L} w_l^2}, \tag{14}$$

where w_l is the gain of the lth loudspeaker radiating sound from a position denoted by the vector $\hat{\mathbf{u}}_\mathbf{l}$ from the loudspeaker's position to the center of the array. The magnitude of the vector, r_E, represents the concentration of source energy in the desired direction, and thus the accuracy of high-frequency localization from that direction. In order to maximize this value, one can apply correcting gains $(g_0, g_1, ..., g_N)$ to each order of spherical harmonics (i.e. to the right side of Eq. 12.) Methods for calculating the correcting gains are given in [3]. This is known as max-r_E decoding.

4.4. Binaural Cue Analysis

The localization accuracy of different decoding schemes can be examined using a model of the auditory system used previously in [15] to evaluate the localization of various stereo recording techniques. The auditory periphery is simulated with Head-Related Transfer Functions (HRTFs). The behavior of the basilar membrane and the hair cells is simulated with a gammatone filter bank with 72 bands and half-wave rectification. ITD analysis is performed using an interaural cross-correlation in each frequency band. ILD analysis is performed using an array of excitation/inhibition (EI) cells. This model allows for the creation of maps that corre-

late ITD and ILD values (expressed in milliseconds and decibels, respectively) with azimuthal directions.

This allows for another way to evaluate the accuracy of ambisonic reproduction - by calculating errors in the ITD and ILD cues rather than errors in the reproduced wavefield, as detailed earlier. First, the spherical harmonic signals are decoded to a virtual 24-channel horizontal loudspeaker array (with equiangular spacing, putting each channel 15 degrees apart in azimuth from its neighbors.) The signals reaching each ear are obtained by convolving the loudspeaker feeds with the appropriate HRTFs and summing over all loudspeakers. Plane waves can then be simulated from a variety of directions, and the ITD and ILD cues can be compared to the natural condition.

From these maps we can then look at average binaural cue error as a function of frequency by averaging over azimuthal angles and also over a number of different HRTF catalogs, as measured in [16]. This is shown in Fig. 4 for 1st through 7th order ambisonic rendering. The dotted lines indicate the average error across all frequency bands. The results are in line with what we would expect from an analysis of the wavefields:

1. Cue errors are less at lower frequencies
2. The higher the order, the higher the frequency at which significant cue errors occur
3. The higher the order, the lower the average cue error across all frequency bands

5. BAND-SPLITTING FILTERS

As discussed in the previous section, it is possible to achieve more perceptually accurate decoding by using basic decoding in lower frequencies and max-r_E decoding in higher frequencies. This necessitates a crossover network to transition between the different decoding schemes. One popular crossover is the Linkwitz-Riley (LR) crossover [17]. A second-order LR crossover is formed from cascading two first-order Butterworth filters, a fourth-order LR crossover from two second-order Butterworth filters, etc. The benefit of the LR crossover is that the low- and high-pass portions are phase-matched, and the magnitude of each portion is -6 dB at the crossover frequency, leading to a unity gain for the sum. The rolloff is equal to 6 dB per octave multiplied by the order.

Depending on the characteristics of the particular systems being used, one might want to decode the signals in three or more bands, requiring multiple crossovers. Putting two crossovers too close to one another in frequency results in a summed signal whose magnitude deviates noticeably from unity gain. These problems can be avoided by placing crossovers far enough apart in frequency so that the magnitudes of the low-pass and high-pass components are equal at a maximum value of -20 dB.

6. PROCESSING MEASURED SOUNDFIELDS FOR PLAYBACK

In this section, four hypothetical spherical microphone arrays are considered: two 16-channel and two 64-channel rigid spherical arrays, each set with radii of 2.5 and 5 cm. These arrays utilize a nearly-uniform sampling scheme, with the positions of the sensors given in [18]. Therefore, the 16-channel arrays are capable of decomposing the spherical harmonics up to third order, the 64-channel arrays up to seventh order.

Fig. 5 shows the White Noise Gain thresholds (the lower line corresponding to a WNG value of -30 dB, the upper to -20 dB) and aliasing frequencies, plotted together with the thresholds for transitioning from basic to max-r_E decoding. (Note that the decoding thresholds are based on the size of a typical human head, and thus are not affected by the properties of the microphone array.) Overlaid on these plots are the decoding schemes chosen based on the principles outlined previously, namely:

1. Do not utilize spherical harmonic components below a WNG of -20 to -30 dB
2. Do not decode above the spherical harmonic aliasing frequency
3. Use max-r_E decoding at frequencies above which the normalized radial error reaches -10 to -20 dB
4. Separate multiple crossovers by at least 20 dB.

As before, where we compared different orders of ambisonics, we can evaluate the accuracy of the ITD and ILD cues of these various decoding schemes (of “mixed" order) across multiple frequency bands, as shown in Fig. 6 (with the plots of 1st and 5th order ambisonics shown for reference). The average errors across all frequency bands shown in Table 1.

Table 1: *Average ITD and ILD cue error and aliasing frequency for each spherical microphone array.*

Array	Avg. ITD error (ms)	Avg. ILD error (dB)	Aliasing Frequency (Hz)
a=2.5 cm, Q=16	0.24	4.3	6551
a=5 cm, Q=16	0.21	4.1	3275
a=2.5 cm, Q=64	0.20	4.1	15285
a=5 cm, Q=64	0.17	3.8	7643

There are two main points that we can gather from Fig. 6. The first is that increasing the number of channels for a spherical microphone array of a given radius is a “win-win” scenario: higher spherical harmonic orders become available, the WNG threshold frequencies for lower orders are pushed lower in frequency, and the aliasing frequency moves higher, yielding a wider frequency range for reconstruction. Thus, we see improvements in both localization accuracy and bandwidth. Of course, increasing the number of channels will increase the cost and complexity of building the array. In addition, going from a 3rd order to a 7th order microphone array does not yield the same gains in localization accuracy as moving from 3rd order to 7th order ambisonic reproduction of a simulated soundfield, as the highest order components are sometimes available for only 1 or 2 octaves before encountering the aliasing frequency.

Increasing the spherical array radius, however, involves a tradeoff. On the positive side, the WNG thresholds move lower, making higher order spherical harmonic components available at lower frequencies. However, the aliasing frequency also moves lower, meaning that we are trading localization accuracy for bandwidth.

7. CONCLUSION

The common basis in spherical harmonics makes ambisonics a natural fit for reproducing soundfields measured with spherical microphone arrays. Incorporating higher order spherical harmonic

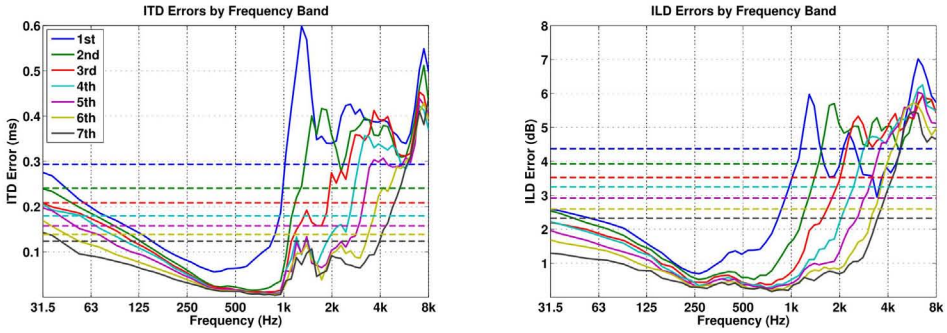


Figure 4: ITD and ILD error by frequency band for 1st through 7th order ambisonics, with average error over all frequency bands shown by the dotted lines.

components offers the opportunity for more precise localization, but at the same time introduces complexities at both the recording stage and the playback stage that need to be dealt with. The goal of this paper is to illuminate the sources of those issues and develop a framework to resolve them, particularly with respect to auditory perception.

8. REFERENCES

[1] Boaz Rafaely, "Plane-wave decomposition of the sound field on a sphere by spherical convolution," *Journal of the Acoustical Society of America*, vol. 116, no. 4, pp. 2149–2157, October 2004.

[2] Zhiyun Li and Ramani Duraiswami, "Flexible and optimal design of spherical microphone arrays for beamforming," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 2, pp. 702–714, February 2007.

[3] Jérôme Daniel, Jean-Bernard Rault, and Jean-Dominique Polack, "Ambisonic encoding of other audio formats for multiple listening conditions," in *Proc. of the 105th Convention of the Audio Eng. Soc.*, San Francisco, California, USA, September 26–29, 1998.

[4] M. A. Poletti, "Three-dimensional surround sound systems based on spherical harmonics," *Journal of the Audio Engineering Society*, vol. 53, no. 11, pp. 1004–1025, November 2005.

[5] Aaron J. Heller, Eric M. Benjamin, and Richard Lee, "A toolkit for the design of ambisonic decoders," in *Linux Audio Conference*, CCRMA, Stanford University, CA, USA, April 12-15 2012.

[6] Earl G. Williams, *Fourier Acoustics: Sound Radiation and Nearfield Acoustical Holography*, Academic Press, 1999.

[7] Philip M. Morse and K. Uno Ingard, *Theoretical Acoustics*, McGraw-Hill, Inc., 1968.

[8] Jens Meyer and Gary W. Elko, "A highly scalable spherical microphone array based on an orthonormal decomposition of the soundfield," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, Orlando, Florida, USA, May 13–17, 2002.

[9] Jens Meyer, Gary W. Elko, and Tony Agnello, "Spherical microphone array for spatial sound recording," in *Proc. of the 115th Convention of the Audio Eng. Soc.*, New York, New York, USA, October 10–13, 2003.

[10] Jens Meyer and Gary W. Elko, *Audio Signal Processing for Next-Generation Multimedia Communication Systems*, chapter 3, "Spherical Microphone Arrays for 3D Sound Recording", pp. 67–89, Kluwer Academic Publishers, Hingham, Massachusetts, USA, 2004.

[11] Boaz Rafaely, Barak Weiss, and Eitan Bachmat, "Spatial aliasing in spherical microphone arrays," *IEEE Transactions on Signal Processing*, vol. 55, no. 3, pp. 1003–1010, March 2007.

[12] Michael A. Gerzon, "Periphony: With-height sound reproduction," *Journal of the Audio Engineering Society*, vol. 21, no. 1, pp. 2–10, January/February 1973.

[13] Darren B. Ward and Thushara D. Abhayapala, "Reproduction of a plane-wave sound field using an array of loudspeakers," *IEEE Transactions on Speech and Audio Processing*, vol. 9, no. 6, pp. 697–707, September 2001.

[14] Michael A. Gerzon, "General metatheory of auditory localisation," in *Proc. of the 92nd Convention of the Audio Eng. Soc.*, Vienna, Austria, March 24-27 1992.

[15] Jonas Braasch, "A binaural model to predict position and extension of spatial images created with standard sound recording techniques," in *Proc. of the 119th Convention of the Audio Eng. Soc.*, New York, NY, USA, October 7–10, 2005.

[16] Jonas Braasch and Klaus Hartung, "Localization in the presence of a disctracter and reverberation in the frontal horizontal plane. i. psychoacoustical data," *Acta Acustica united with Acustica*, vol. 88, no. 6, pp. 942–955, November/December 2002.

[17] Siegfried H. Linkwitz, "Active crossover networks for noncoincident drivers," *Journal of the Audio Engineering Society*, vol. 24, no. 1, pp. 2–8, January/February 1976.

[18] J. Fliege and U. Maier, "The distribution of points on the sphere and corresponding cubature formulae," *IMA Journal of Numerical Analysis*, vol. 19, no. 2, pp. 317–334, 1999.

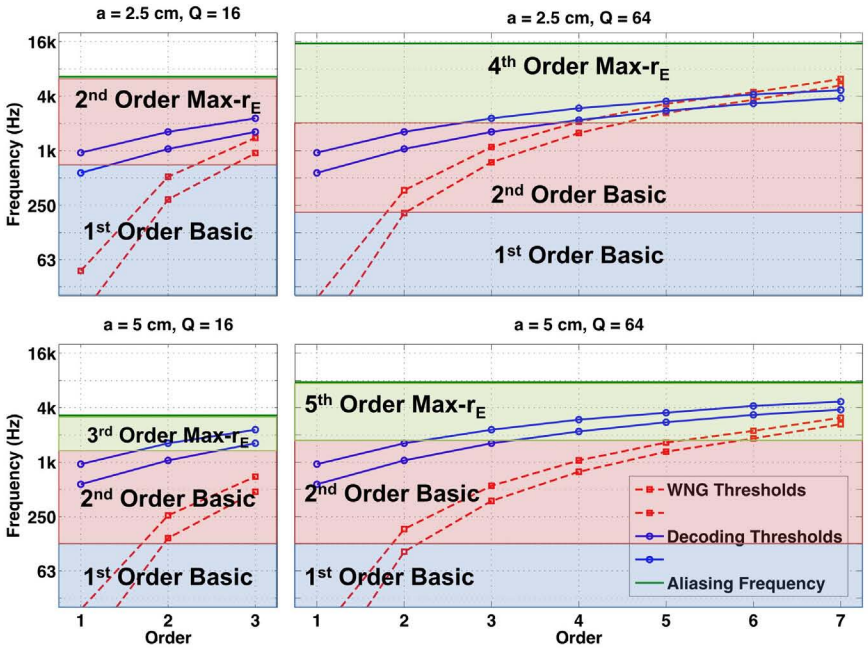


Figure 5: WNG thresholds and aliasing frequencies for 4 spherical arrays, plotted together with basic/max-r_E ambisonic decoding thresholds, overlaid with decoding schemes.

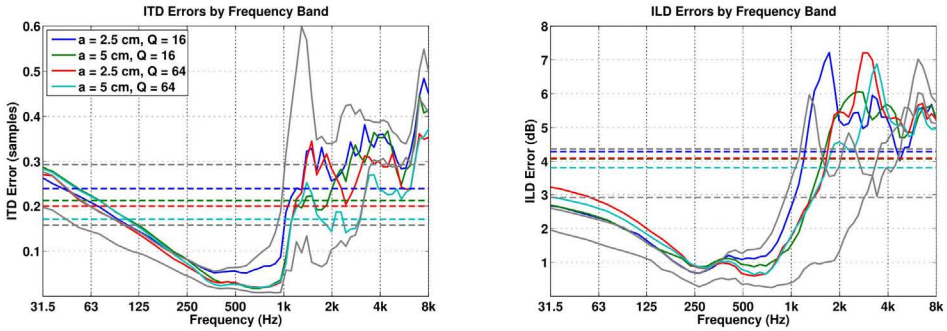


Figure 6: ITD and ILD errors for the 4 different decoding schemes, with the errors for 1st and 5th order ambisonic decoding shown for reference.

Proc. of the EAA Joint Symposium on Auralization and Ambisonics, Berlin, Germany, 3-5 April 2014

THE IMPACT OF THE WHITE NOISE GAIN (WNG) OF A VIRTUAL ARTIFICIAL HEAD ON THE APPRAISAL OF BINAURAL SOUND REPRODUCTION

Eugen Rasumow, Matthias Blau, Martin Hansen, *

Institute of hearing technology and audiology
Jade University of Applied Sciences
Oldenburg, Germany
eugen.rasumow@jade-hs.de

Simon Doclo, Steven van de Par, Volker Mellert

Institute of Physics
Carl-von-Ossietzky University
Oldenburg, Germany

Dirk Püschel

Akustik Technologie Göttingen
Göttingen, Germany

ABSTRACT

As an individualized alternative to traditional artificial heads, individual head-related transfer functions (HRTFs) can be synthesized with a microphone array and digital filtering. This strategy is referred to as "virtual artificial head" (VAH). The VAH filter coefficients are calculated by incorporating regularization to account for small errors in the characteristics and/or the position of the microphones. A common way to increase robustness is to impose a so-called white noise gain (WNG) constraint. The higher the WNG, the more robust the HRTF synthesis will be. On the other hand, this comes at the cost of decreasing the synthesis accuracy for the given sample of the HRTF set in question. Thus, a compromise between robustness and accuracy must be found, which furthermore depends on the used setup (sensor noise, mechanical stability etc.). In this study, different WNG are evaluated perceptually by four expert listeners for two different microphone arrays. The aim of the study is to find microphone array-dependent WNG regions which result in appropriate perceptual performances. It turns out that the perceptually optimal WNG varies with the microphone array, depending on the sensor noise and mechanical stability but also on the individual HRTFs and preferences. These results may be used to optimize VAH regularization strategies with respect to microphone characteristics, in particular self noise and stability.

1. INTRODUCTION

In order to take into account spatial cues within a binaural reproduction, the use of so-called artificial heads, which are a replica of real human heads and pinnae, is common practice today. By this means the signals at the ears receive characteristic spatial information, which encompasses interaural time and level difference cues, but also spectral cues due to the shape of the pinna, for instance. Disadvantageously, artificial heads are inherently bound to non-individual (average) anthropometric geometries and are most often implemented as bulky devices. Alternatively, the individual frequency-dependent directivity patterns of a human head (HRTFs) can be synthesized with a microphone array and digital filtering (cf, [1], [2], [3], [4] and [5]), which will be referred to as a virtual artificial head (VAH). A VAH is more flexible than real artificial heads, since, e.g., the filters can be adjusted post-hoc to match any individual sets of HRTFs. In contrast to approaches in the spherical harmonics domain (i.e. applying spherical harmonics decomposition, optimization and re-synthesis, cf. [3] and [6]), the VAH re-synthesis in this study is optimized in the frequency domain for discrete directions in the horizontal plane only, assuming the intermediate directions to be inherently interpolated by the VAH. One advantage of this approach is that much fewer microphones are needed in comparison to e.g. spherical harmonics based approaches (cf. [7] and [8]). The individual filter coefficients can be calculated by optimizing various cost functions, where a least square cost function is known to yield appropriate perceptual results (cf. [5]) and is thus used in this study (cf. section 2). The robustness of the filter coefficients is usually assured by imposing a constraint on the so-called white noise gain (WNG), in order to consider small deviations of the microphone characteristics and/or positions (cf. [4]). By doing so, the robustness of the filter coefficients increases with higher WNG while the accuracy decreases at the same time for a given HRTF set and vice versa (cf. Figure 1). Thus, it seems reasonable to find a compromise in the regularization, where the perceptual appraisal of a HRTF re-synthesis using the VAH is assessed best as a function of the WNG. Two microphone arrays (cf. Figure 2) were applied in this study. These arrays enabled the use of measured steering vectors (as opposed to the application of analytical steering vectors in cf. [3], [4] or [6]) and to re-synthesize individual ear signals by individually recalculating pre-recorded signals.

2. REGULARIZED LEAST SQUARES COST FUNCTION

Consider the desired directivity pattern $\mathrm{D}(\omega, \Theta)$ as a function of frequency ω and discrete azimuthal angles Θ, as well as the $\mathrm{N}\times 1$ steering vector $\mathbf{d}(\omega, \Theta)$ which represent the frequency- and direction-dependent transfer functions between the source and the N microphones. Then the re-synthesized directivity pattern of the VAH $\mathrm{H}(\omega, \Theta)$ for one particular set of steering vectors $\mathbf{d}(\omega, \Theta)$

* Author to whom correspondence should be addressed. Electronic mail: eugen.rasumow@jade-hs.de

can be expressed as[1]

$$\mathrm{H}(\omega,\Theta) = \mathbf{w}^{\mathbf{H}}(\omega)\mathbf{d}(\omega,\Theta). \tag{1}$$

Here, the N×1 vector $\mathbf{w}(\omega)$ contains the complex-valued filter coefficients for each microphone per frequency ω and a given set of steering vectors $\mathbf{d}(\omega,\Theta)$.
In order to calculate the filter coefficients $\mathbf{w}(\omega)$ for the steering vectors $\mathbf{d}(\omega,\Theta)$, one may employ a narrowband least squares cost function $\mathrm{J_{LS}}$, being the sum over P directions of the squared absolute differences between $\mathrm{H}(\omega,\Theta)$ and $\mathrm{D}(\omega,\Theta)$ that is to be minimized, i.e.

$$\mathrm{J_{LS}}(\mathbf{w}(\omega)) = \sum_{\Theta=1}^{\mathrm{P}} \left| \mathbf{w}^{\mathbf{H}}(\omega)\mathbf{d}(\omega,\Theta) - \mathrm{D}(\omega,\Theta) \right|^2 . \tag{2}$$

In this study, filters were optimized to represent individual HRTFs measured in the horizontal plane with an equidistant angular spacing of $\Delta\Theta = 15°$, resulting in P = 24 directions. A straightforward minimization of Eq. 2, however, may result in non robust filter coefficients $\mathbf{w}(\omega)$, where already small errors of the microphone positions and/or characteristics may cause huge errors of the re-synthesized directivity patterns (cf. [4] and [9]) and which may lead to a not desirable amplification of spatially uncorrelated noise at the microphones. More robust filter coefficients can be obtained by imposing a constraint on the derived filter coefficients. To this end, we propose a modified definition of the white noise gain ($\mathrm{WNG_m}$), given as

$$\mathrm{WNG_m}(\omega) = 10 \cdot \log_{10}\left(\frac{\mathbf{w}^{\mathbf{H}}(\omega)\mathbf{Q_m}(\omega)\ \mathbf{w}(\omega)}{\mathbf{w}^{\mathbf{H}}(\omega)\mathbb{I}_{\mathrm{N}}\ \mathbf{w}(\omega)} \right), \text{ with}$$

$$\mathbf{Q_m}(\omega) = \frac{1}{\mathrm{P}} \sum_{\Theta=1}^{\mathrm{P}} \mathbf{d}(\omega,\Theta)\mathbf{d}^{\mathbf{H}}(\omega,\Theta) \tag{3}$$

and $\mathbb{I}_{\mathrm{N}}$ being the N×N-dimensional unity matrix. By doing so, $\mathrm{WNG_m}(\omega)$ relates the mean array gain in the measured acoustic field (determined by $\mathbf{Q_m}(\omega)$ and $\mathbf{w}(\omega)$) to the inner product of the filter coefficients, i.e. to the array gain for spatially uncorrelated noise at the microphones (cf. [10]). Usually, regarding beamforming applications the WNG is given for a certain direction (discrete steering direction Θ_0) only (cf. [11],[12] and [5]), whereas the $\mathrm{WNG_m}$ in Eq. 3 may be referred to as the mean WNG over all considered directions Θ. This modification of the WNG was applied since a direction-dependent constraint (as is realized in the classical WNG) would consequently yield a direction-dependent regularization, which is not desirable for a VAH re-synthesis. Hence, the mean $\mathrm{WNG_m}$ incorporating all associated directions is introduced in this study (Eq. 3). Positive $\mathrm{WNG_m}$ represent an attenuation of spatially uncorrelated noise, whereas negative $\mathrm{WNG_m}$ represent an amplification ([11]) relative to the mean array gain in the measured acoustic field. We suggest to apply the constraint $\mathrm{WNG_m}(\omega) \geq \beta$ for regularization, where the gain β (in dB) has to be chosen manually according to the expected error of the steering vectors (cf. [4]). The combination of the least squares cost function from Eq. 2 with the constraint incorporating Eq. 3 results in the cost function

$$\mathrm{J_{LS}}_{\rho}(\mathbf{w}(\omega)) = \sum_{\Theta=1}^{\mathrm{P}} \left| \mathbf{w}^{\mathbf{H}}(\omega)\mathbf{d}(\omega,\Theta) - \mathrm{D}(\omega,\Theta) \right|^2 + \mu\left(\left(\mathbf{w}^{\mathbf{H}}(\omega)\mathbb{I}_{\mathrm{N}}\mathbf{w}(\omega) \right) - \frac{1}{\beta_{\mathrm{pow}}} \left(\mathbf{w}^{\mathbf{H}}(\omega)\mathbf{Q_m}(\omega)\ \mathbf{w}(\omega) \right) \right), \tag{4}$$

where μ represents the Lagrange multiplier and $\beta_{\mathrm{pow}} = 10^{\frac{\beta}{10}}$.
The closed form solution of $\mathrm{J_{LS}}_{\rho}(\mathbf{w}(\omega))$, yielding the regularized filter coefficients $\mathbf{w}(\omega)$, is given by

$$\mathbf{w}(\omega) = \left(\mathbf{Q}(\omega) + \mu\left(\mathbb{I}_{\mathrm{N}} - \frac{1}{\beta_{\mathrm{pow}}} \cdot \mathbf{Q_m}(\omega) \right) \right)^{-1} \cdot \mathbf{a}(\omega)\ , \quad \text{with} \tag{5}$$

$$\mathbf{Q}(\omega) = \sum_{\Theta=1}^{\mathrm{P}} \mathbf{d}(\omega,\Theta)\mathbf{d}^{\mathbf{H}}(\omega,\Theta) \ \text{ and} \tag{6}$$

$$\mathbf{a}(\omega) = \sum_{\Theta=1}^{\mathrm{P}} \mathbf{d}(\omega,\Theta)\ \mathrm{D}^*(\omega,\Theta)\ . \tag{7}$$

While the least squares solution of the cost function in Eq. 2 is quite well known in literature (cf. [9], [5]), the regularization term in Eq. 5 differs from usual regularization strategies, as for instance known from diagonal loading (cf. [13]), Tikhonov-regularization or similar regularization approaches (cf. [14]). The main difference lies in the dependence of the regularization on the applied steering vectors ($\mathbf{Q_m}(\omega)$) and the desired $\mathrm{WNG_m}$ β. However, the presented regularization approaches the diagonal loading or Tikhonov-regularization for very large β_{pow} (i.e., for the most stringent regularization possible).
The optimal μ to satisfy the desired WNG-constraint was chosen iteratively. Analogous to the procedure in [5], μ was increased in steps of $\Delta\mu = \frac{1}{100}$ for each ω until $\mathrm{WNG_m}(\omega,\mu) \geq \beta$ or $\mu_{\mathrm{max}} = 100$ were reached (if existent at all, this only occurred at very high frequencies).

2.1. Influence of the WNG-constraint on the VAH re-syntheses

The accuracy of the VAH re-syntheses depends on the desired HRTFs, the number of microphones, the topology of the microphone array, the cost function and also the applied Lagrangian

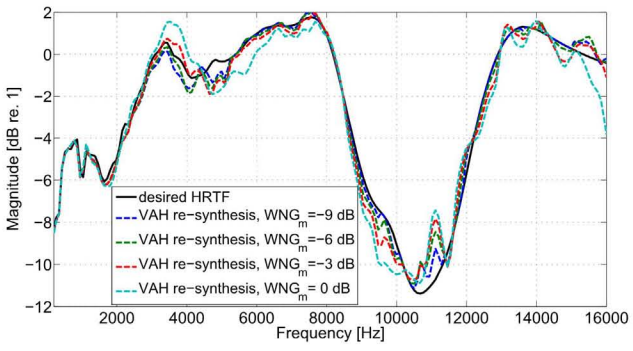


Figure 1: *Magnitude of the desired HRTF ($\Theta = 90°$) for the left ear of subject* $\mathrm{S_1}$ *(black line) and VAH re-syntheses with various* $\mathrm{WNG_m}$ *(dashed lines) for* $\mathrm{array_2}$ *as a function of frequency.*

[1] In the following $x^{\mathbf{H}}$ denotes the Hermitian transpose of x and x^* denotes the complex conjugate of x.

multiplier μ (cf. Eq. 5). In general, the desired $\mathrm{WNG_m}$ is approached by gradually increasing μ. This in turn will cause increasing deviations of the re-syntheses from the desired HRTF. The magnitude of the resulting μ is primarily determined by the desired $\mathrm{WNG_m}$ β. Thus, the regularization yielding a desired $\mathrm{WNG_m}$ unavoidably causes distortions of the VAH re-syntheses which may vary individually with the desired HRTFs and steering vectors. This aspect is exemplarily depicted in Figure 1. On the other hand, higher $\mathrm{WNG_m}$ are associated with more robustness regarding small changes of the microphone characteristics and/or with a lower amplification of spatially uncorrelated noise at the microphones.

3. MICROPHONE ARRAYS USED

The main goal of this study is to investigate the perceptually optimal $\mathrm{WNG_m}$ for different subjects, using different microphone arrays. For this reason, the perceptual evaluation was made with recordings using two open planar microphone arrays incorporating different kinds of microphones and support structures but the same number of microphones and an identical topology which was chosen according to [4]. The advantage of using open planar arrays over rigid spheres or the like is the opportunity to realize various two-dimensional inter-microphone distances. By this means, a mathematically motivated microphone topology according to [4] was chosen, which is assumed to yield appropriate results regarding the accuracy and robustness of the re-syntheses.
The first microphone array ($\mathrm{array_1}$, left panel in Figure 2) consisted of 24 Sennheiser KE 4-211-2 microphones. The individual microphones were mounted on a wooden plate using a solid wire construction. Together with analog preamplifiers the sensor noise of each single microphone signal was approximately 35 dB(A). No absorbent material was used for the support structure of $\mathrm{array_1}$.

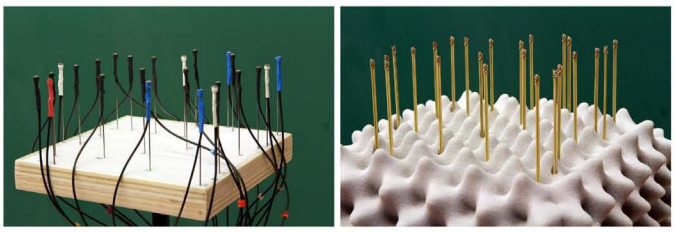


Figure 2: *Two used microphone arrays with 24 KE-4 microphones ($\mathrm{array_1}$, left) and 24 sensors composed of 48 MEMS microphones ($\mathrm{array_2}$, right) with the same planar microphone topology according to [4].*

For the second array ($\mathrm{array_2}$), micro-electromechanical system (MEMS) microphones (Analog Devices ADMP 504 Ultralow Noise Microphone) were used in an custom-made electrical circuit. Here, each sensor is composed of two MEMS microphones. A composed sensor yielded a sensor noise of approximately 27 dB(A), which is quite low for this kind of microphones. The directivity of such a composed sensor can be assumed to be negligible for frequencies of interest (i.e. $f \lesssim 16$ kHz). For $\mathrm{array_2}$, 24 of these sensors (consisting of 48 MEMS microphones) were mounted on a printed circuit board (cf. right panel in Figure 2) with the same topology as for $\mathrm{array_1}$. In order to reduce effects of standing waves between the sensors and the board, $\mathrm{array_2}$ is covered with absorbent material.

4. EXPERIMENTAL PROCEDURE

4.1. Material

Prior to the experiment, individual HRTFs and headphone (AKG K-240 Studio) transfer functions (HPTFs) were measured for four subjects using the blocked ear method according to [15]. For measuring the HPTFs, subjects were instructed to reposition the headphone ten times to various realistic carrying positions which successively yielded ten different individual HPTFs. The individual HPTF resulting in the smallest dynamic range of its magnitude for frequencies 300 Hz $\leq f \leq$ 16000 Hz was inverted in the frequency domain and transformed into the time domain. The HRTFs as well as the inverse HPTFs were implemented as finite impulse response (FIR) filters with a filter length of 256 taps, corresponding to ≈ 5.8 ms at a sampling frequency of $f_s = 44100$ Hz. This filter length was chosen to incorporate all aspects associated with an appropriate binaural reproduction (cf. [16]). The individual HRTFs as well as the steering vectors $\mathbf{d}(\omega, \Theta)$ for the two microphone arrays were measured in the horizontal plane with an angular spacing of 15°. All HRTFs were smoothed in the frequency and spatial domain prior to the VAH re-syntheses according to the perceptual limits derived in [17]. Moreover, the associated impulse responses of all measured steering vectors $\mathbf{d}(\omega, \Theta)$ were also truncated to a filter length of 256 taps in order to achieve smoother transfer functions.

4.2. Test stimulus

As to cover a wide frequency range and simultaneously to include temporal cues, the test stimulus for perceptual evaluation consisted of 3 short bursts of pink noise filtered with an eighth order bandpass with the cutoff frequencies of $f_{\mathrm{low}} = 300$ Hz and $f_{\mathrm{hi}} = 16000$ Hz. The lower bandwidth limitation of the test stimulus f_{low} was chosen due to the limits of the loudspeakers used. However, since the influence of varying the $\mathrm{WNG_m}$ is primarily evident for frequencies $f \geq 3$ kHz (cf. Figure 1) it seems reasonable to assume that this limitation does not have a significant influence on the perceptual evaluations. Each noise burst lasted $\frac{1}{3}$ s with 0.01 s onset-offset ramps followed by silence of $\frac{1}{6}$ s. This test stimulus was intended to facilitate the evaluation of spectral deviations, temporal dispersion but also the influence of the sensor noise. The presented stimuli were calibrated with a G.R.A.S. type 43AA artificial ear to have 70 dB SPL for the frontal direction $\Theta = 0^\circ$.

4.3. Methods

A listening test was carried out with four experienced listeners (two of them are authors of this article). The subjects were instructed to rate four different aspects (localization, sensor noise, overall performance and spectral coloration, cf. section 4.3.1) of a test presentation with respect to the reference presentation (binaural reproduction with original individual HRTFs and HPTFs). The quality of the reference setting (representing desirable re-syntheses) has a major effect on the evaluations. Thus it needed to be assured that the individual binaural reproductions incorporated all essential individual spatial characteristics. For this reason, the individual binaural reproductions used in the reference setting were played to the subjects before the experimental procedure in a preliminary listening test. All subjects were able to perceive the presented stimuli outside the head and correctly assigned the corresponding directions in the horizontal plane.

Prior to the listening tests, the steering vectors were measured and the test stimuli were recorded using the two microphone arrays (cf. section 3) in an anechoic chamber. Furthermore, the individual VAH filters were optimized to re-synthesize the individual HRTFs in the horizontal plane with an angular spacing of $\Delta\Theta = 15°$. In the test condition, the sum of the filtered stimuli (representing the re-synthesized ear signals, cf. Eq.1) was also filtered with the inverse HPTF filters (same procedure as in the reference setting) and played to the subject via headphones. In both conditions, the stimuli were played back in an infinite loop with the possibility to switch between the reference- and test condition or to stop the playback. To limit the number of experiments to a manageable amount, three directions in the horizontal plane were chosen for evaluation with azimuth angles $\Theta = 0°$ (front), $\Theta = 90°$ (left) and $\Theta = 225°$ (back right) and the $\mathrm{WNG_m}$ was one of $\mathrm{WNG_m}(\omega)$ = -9 dB, -6 dB, -3 dB or 0 dB for all ω. These preselected $\mathrm{WNG_m}$ were assumed to roughly cover the area with the best suited $\mathrm{WNG_m}$ based on previous preliminary tests.

The three tested azimuthal directions Θ, the two microphone arrays as well as the four $\mathrm{WNG_m}$ were varied in randomized order within one experimental run with three random presentations (retest) for each condition. The true identities of the signals in the reference and test setting were hidden to the subjects. In sum, 216 conditions (presented signal pairs) were evaluated by each subject, whereas one of the tested parameters (impact of various calibration strategies) was eliminated from the analysis in this article in hindsight. Hence, 3 directions $\times$ 2 arrays $\times$ 3 presentations $\times$ 4 $\mathrm{WNG_m}$ = 72 individual evaluations (of a total of originally 216 individually gathered evaluations) will be analyzed and discussed in section 5 and 6. Within each condition, subjects were able to switch between the reference and the test setting arbitrarily. The entire experiment was performed applying an English category scale, ranging between *bad*, *poor*, *fair*, *good* and *excellent* with four intermediate undeclared steps (cf. [5]). Each session lasted approximately 120-180 minutes, where subjects were able to subdivide the session arbitrarily and to do as many breaks as they wanted. Prior to the evaluation each subject had time for familiarization with the various reference and test conditions.

4.3.1. Assessed aspects

The subjects were instructed to evaluate the quality of the test setting with respect to reference setting for four chosen aspects which are assumed to be significant for appropriate VAH re-syntheses:

- localization: The evaluation of localization incorporated the perceived angle of incidence (azimuth and elevation) and the perceived distance in combination.
- sensor noise: Subjects were instructed to evaluate the perceived sensor noise which was primarily apparent in the temporal pauses of the test stimulus.
- overall performance: The evaluation of the perceived overall performance incorporated all feasible aspects depending on the taste and preferences of the individual subject.
- spectral coloration: Subjects were instructed to evaluate the perceived spectral coloration without evaluating the potential deviations of localization or other cues.

5. RESULTS AND DISCUSSION - PERCEPTUAL EVALUATION

The mean and the standard deviations (over three randomized presentations) of all individual evaluations are depicted in Figure 3 as functions of the $\mathrm{WNG_m}$ on the x-axis with the assessed aspects separated in rows, the directions Θ separated in columns and the color indicating the subjects. The average performance (means and standard deviations over subject) is depicted in Figure 4, with the color indicating the assessed aspects (see legend).
In general, the perceptual evaluations and their variation within repeated trials in Figure 3 (standard deviation depicted as error bars) seem to depend on the direction of incidence Θ and the used microphone array, but as well on the subject. This is an effect of individual preferences with individual internal scales and was to be expected according to analogous studies (cf. [5]). In order to analyze potential preferences regarding the $\mathrm{WNG_m}$ for the application of a VAH, primarily the relative tendencies of intra- and inter-individual perceptual evaluations depending on the $\mathrm{WNG_m}$ are focused on.

Table 1: p-values (rounded to 3 digits) according to the Friedman test regarding localization, overall performance, sensor noise and coloration for the three tested directions separately. p-values indicating significantly different evaluations when varying the $\mathrm{WNG_m}$ ($p \leq \frac{0.05}{24} = 0.0021$) are depicted as bold numbers.

localization	array_1	array_2	overall	array_1	array_2
$\Theta = 0°$	0.164	0.445	$\Theta = 0°$	0.341	0.081
$\Theta = 90°$	0.004	0.006	$\Theta = 90°$	**0.000**	0.129
$\Theta = 225°$	0.147	0.933	$\Theta = 225°$	0.109	0.188
sensor noise	array_1	array_2	**coloration**	array_1	array_2
$\Theta = 0°$	0.004	0.049	$\Theta = 0°$	0.035	0.578
$\Theta = 90°$	**0.000**	0.340	$\Theta = 90°$	**0.000**	0.827
$\Theta = 225°$	**0.000**	0.079	$\Theta = 225°$	0.015	0.319

Although means and standard deviations were used for illustrating the evaluations in Figs. 3 and 4 (for increased clarity), a non parametric statistical test was applied. The Friedman test was applied to analyze whether the evaluations for at least one of the tested $\mathrm{WNG_m}$ (for a fixed direction, array and assessed aspect) was considerably different than the evaluations for the other $\mathrm{WNG_m}$. A sufficiently small p-value indicated an effect of the $\mathrm{WNG_m}$ on the evaluations. The p-values for the assessed aspects (separate boxes), the applied arrays (columns) and directions (rows) are given in Table 1. The p-values for conditions indicating a significant effect of the $\mathrm{WNG_m}$ on the perceptual evaluations (considering the Bonferroni correction for 24 repeated tests, a p-value of $p \leq \frac{0.05}{24}$ is assumed to indicate a significant effect of the $\mathrm{WNG_m}$) are depicted as bold numbers. However, due to the rather small number of subjects and the presumably low test power, the p-values in Table 1 may primarily be used to highlight tendencies of all evaluations for fixed conditions without postulating any statistical (in)significances for the effect of the $\mathrm{WNG_m}$.
In sum, it emerges that the tested $\mathrm{WNG_m}$ mainly seem to have an effect on the evaluations for array_1 with regard to sensor noise and coloration. The evaluations regarding localization seem primarily to be affected by the $\mathrm{WNG_m}$ for $\Theta = 90°$ and both arrays. The evaluations regarding the overall performance seem to be affected by the $\mathrm{WNG_m}$ mainly for array_1 and $\Theta = 90°$.

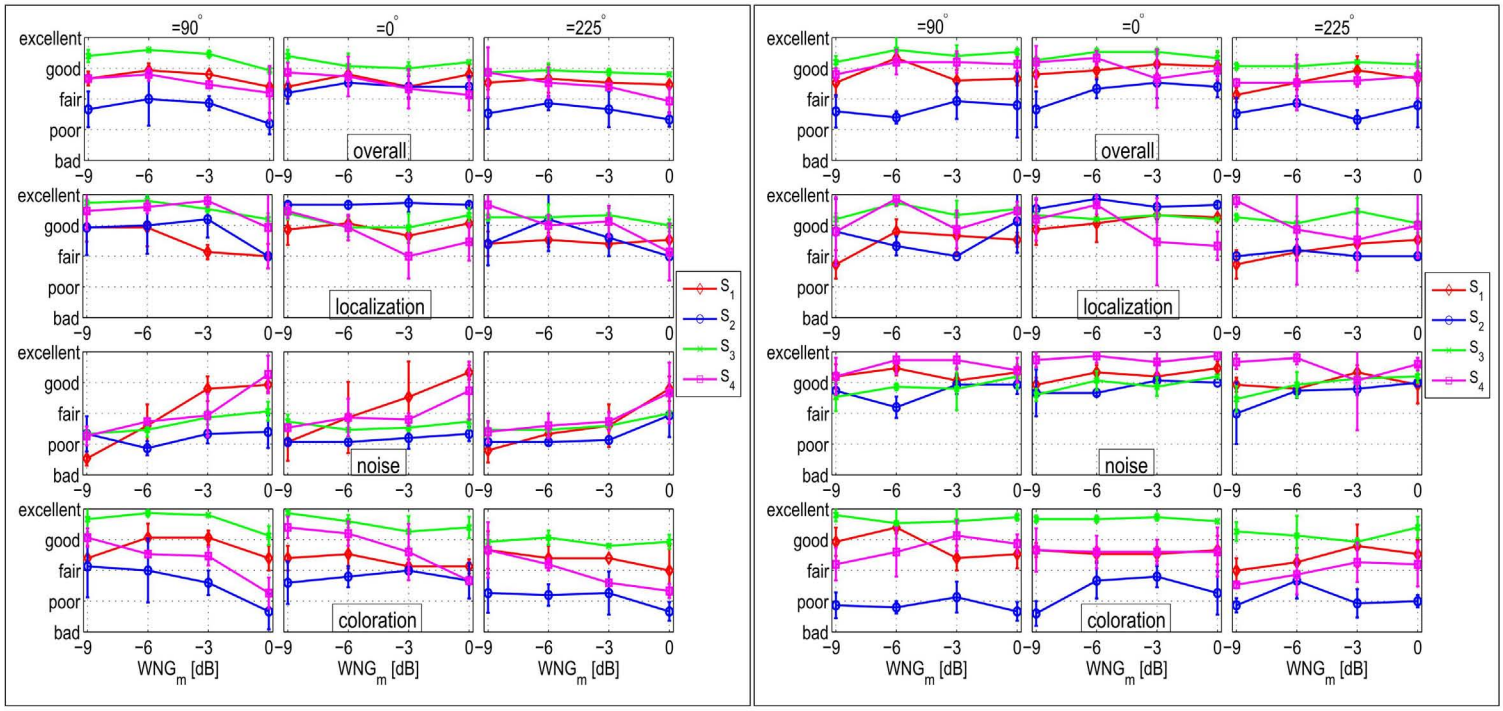


Figure 3: *Perceptual evaluations for* array_1 *(left block) and* array_2 *(right block). The aspects of evaluation are aligned in separate rows (first row: overall performance, second row: localization, third row: sensor noise and fourth row: spectral coloration) and the direction of arrival* Θ *is aligned in three columns (*$\Theta = 90°$ *in the left column,* $\Theta = 0°$ *in the middle column and* $\Theta = 225°$ *in the right column). The individual evaluations (mean and standard deviation over three randomized presentations) are depicted as a function of the* $\mathrm{WNG_m}$ *in dB. The colors and markers indicate the four subjects (*S_1, S_2, S_3 *and* S_4*).*

5.1. Localization

In general, all subjects concordantly reported the localization in the horizontal plane to be re-synthesized well by the VAH. However, the aspect localization was also used to evaluate the perceived distance of the sound source (cf. section 4.3.1). The perception of distance may vary noticeably when interaural level differences from lateral directions are not re-synthesized accurately. This may be a possible explanation for the better evaluations for $\Theta = 0°$, which is especially evident for subject S_1 and S_2 (cf. Figure 3).

For subject S_3, the evaluations with regard to localization vary hardly with the tested $\mathrm{WNG_m}$ nor with the array. The p-values from Table 1 indicate the most notable effect of the $\mathrm{WNG_m}$ on the evaluations with regard to localization for $\Theta = 90°$ with both arrays. This aspect is also apparent in the averaged evaluations (cf. Figure 4) for array_1, where the evaluations decrease for higher $\mathrm{WNG_m}$. However, there does not seem to be such an unambiguous tendency for the evaluations with array_2 and $\Theta = 90°$. Moreover, the averaged evaluations seem also to decrease slightly with increasing $\mathrm{WNG_m}$ for $\Theta = 225°$ and array_1. This slight effect is concordantly associated with a relatively higher p-value from the Friedman test (p=0.147), as well indicating a less notable effect of the tested $\mathrm{WNG_m}$.

In sum, the evaluations of localization seems to decrease with higher $\mathrm{WNG_m}$ using array_1 and are approximately constant or do not vary in a clearly interpretable way for array_2.

5.2. Sensor noise

The evaluations with regard to the perceived sensor noise for array_1 are considerably different from the evaluations for array_2. Especially for lower $\mathrm{WNG_m}$ ($\mathrm{WNG_m} \leq -3$ dB), the sensor noise for array_1 is evaluated worse compared to the evaluations for array_2. The evaluations improve with increasing $\mathrm{WNG_m}$, especially for subjects S_1 and S_4 where the evaluations for $\mathrm{WNG_m}$=0 dB and array_1 are approximately in the range of the evaluations for array_2. The evaluations for array_2 vary much less with the $\mathrm{WNG_m}$, resulting for subjects S_1 and S_4 in variations of approximately the amount of their standard deviations (over randomized presentations). This effect is also represented by the associated p-values, with relatively small p-values ($\mathrm{p} \leq 0.004$) for all directions Θ and array_1 and rather high p-values ($\mathrm{p} \geq 0.049$) for all directions Θ and array_2. On the other hand, there also seems to be a slight trend towards better evaluations for higher $\mathrm{WNG_m}$ with array_2, with the worst evaluations for the lowest $\mathrm{WNG_m}$ of -9 dB (in the averaged evaluations in Figure 4 as well as for subject S_2 and S_3 and $\Theta = 225°$ in Figure 3). This indicates that sensor noise is not negligible for all subjects even with array_2. However, the averaged evaluations in Fig. 4 as well as the associated p-values in Table 1 indicate that the gathered evaluations vary much less with the tested $\mathrm{WNG_m}$ when using array_2 compared to array_1.

In sum, the perceptually optimal $\mathrm{WNG_m}$ with regard to sensor noise seems to vary with the used microphone array and its inherent sensor noise. The evaluations of the sensor noise (if detectable) seem generally to enhance with higher $\mathrm{WNG_m}$, which was to be expected.

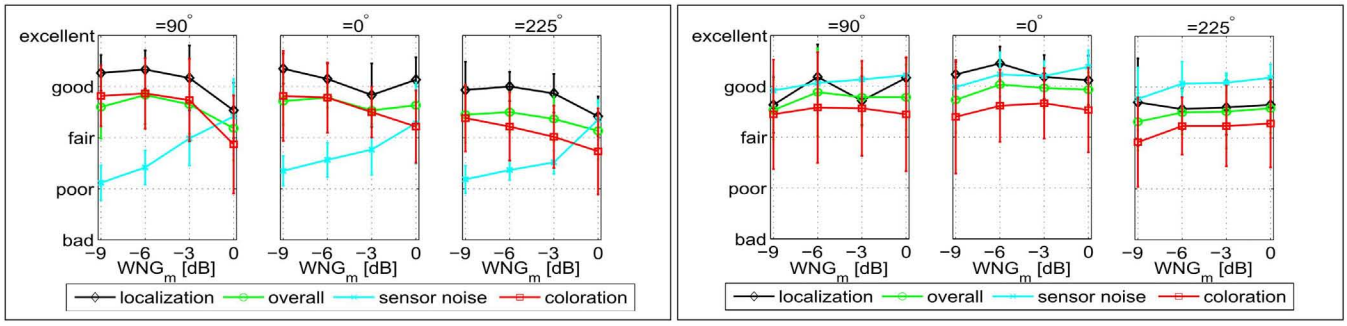


Figure 4: Perceptual evaluations averaged over all subjects for the array_1 (left block) and array_2 (right block) are depicted as the mean and the standard deviation for the four aspects to be evaluated (localization, overall performance, sensor noise and coloration).

5.3. Overall performance

The largest variations of the evaluations with regard to overall performance can be observed across different subjects, while the evaluations remain rather constant over different $\mathrm{WNG_m}$, especially for subject $\mathrm{S_3}$ with both microphone arrays. However, there seems to be a slight trend to worse evaluations for higher $\mathrm{WNG_m}$ using array_1 (cf. $\Theta = 90°$ and $\Theta = 225°$) as well as for the lowest $\mathrm{WNG_m}$ of -9 dB (presumably due to the more disturbing sensor noise). This trend is also apparent from the averaged performance using array_1 in Figure 4, with the Friedman test indicating the largest effect of the $\mathrm{WNG_m}$ for $\Theta = 90°$.

The evaluations vary less clearly with the $\mathrm{WNG_m}$ for array_2. There, the best evaluations were mostly observed at higher $\mathrm{WNG_m}$ (cf. $\mathrm{S_1}$, $\Theta = 225°$ and $\mathrm{S_2}$, $\Theta = 0°$) and worsened slightly for the lowest $\mathrm{WNG_m}$ (cf. Figure 4). In general, the evaluations with regard to overall performance seem to be correlated to the evaluations with regard to spectral coloration (cf. section 5.4), again emphasizing the relevance of spectral coloration for the evaluation of a binaural re-synthesis with respect to a reference condition. Furthermore, comparing the averaged evaluations of the overall performance for both microphone arrays (cf. Figure 4) at higher $\mathrm{WNG_m}$, the evaluations seem better for array_2 compared to array_1. This aspect is assumed to be a consequence of the lower inherent sensor noise of array_2: Typically, the Lagrangian multiplier μ is lower for lower desired $\mathrm{WNG_m}$. To achieve a desired

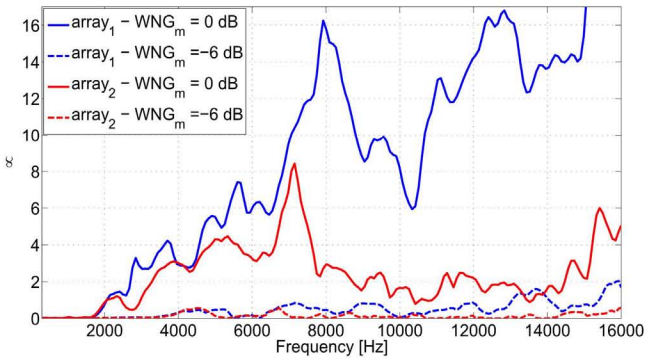


Figure 5: *Exemplary course of the Lagrangian multiplier μ (cf. Eq. 5) for* array_1 *and* array_2 *(blue and red lines, respectively) and* $\mathrm{WNG_m}$ *of 0 dB and -6 dB (solid and dashed lines, respectively) as a function of frequency of the left-ear re-synthesis for* $\mathrm{S_1}$.

$\mathrm{WNG_m}$, the required μ is usually lower (empirical observation) for array_2 compared to array_1, cf. Figure 5. Although not shown here, this tendency has also been observed for the other subjects and $\mathrm{WNG_m}$. A possible explanation could be that μ needs to be enlarged more in order to counteract the higher inherent sensor noise of array_1 (resulting in larger random errors on the measured steering vectors) in comparison to array_2. Considering that the accuracy of a re-synthesis decreases with larger μ, the higher inherent sensor noise of array_1 may therefore be a reasonable explanation for a worse accuracy of the re-syntheses and subsequently for the worse evaluations at $\mathrm{WNG_m} \gtrsim -3$ dB.

In sum, the evaluations with regard to overall performance seem best for $\mathrm{WNG_m}$=-6 dB and $\mathrm{WNG_m}$=-3 dB when using array_1 and for $\mathrm{WNG_m} \geq$=-6 dB when using array_2.

5.4. Spectral coloration

The evaluations with regard to spectral coloration seem to differ considerably for the four subjects. This phenomenon may be partly explained by the fact that the perception and evaluation of spectral coloration is influenced by the perceived localization and the interaction with the perceived sensor noise. This may introduce a certain degree of interpretation to assess this aspect. Furthermore, subjects have individual internal scales and assess individually. This is primarily evident when comparing the evaluations of subject $\mathrm{S_2}$ and $\mathrm{S_3}$, for instance. The evaluations of subject $\mathrm{S_3}$ vary roughly between good and excellent while the evaluations of subject $\mathrm{S_2}$ vary roughly between fair and poor, representing the most negative evaluations of this study.

In general, slightly better evaluations are evident for the frontal direction $\Theta = 0°$ compared with the lateral directions. The averaged evaluations in Figure 4 as well as the p-values in Table 1 indicate that the evaluations for array_1 vary considerably across the tested $\mathrm{WNG_m}$ for all tested directions Θ with decreasing averaged evaluations for higher $\mathrm{WNG_m}$ in Figure 4. This tendency does, however, not hold for array_2, with its p-values being relatively high ($p \geq 0.319$) for all directions. This array-dependent difference of evaluations may be explained by the differently sized Lagrangian multipliers μ for the two applied arrays (cf. Figure 5 and the discussion in section 5.3).

In sum, the evaluations of the perceived spectral coloration seem to vary with subjects and also with the used microphone arrays. Higher $\mathrm{WNG_m}$ seem to distort the perception of spectral coloration for array_1. On the other hand, the evaluations with regard to spectral coloration do not seem to vary considerably with the tested $\mathrm{WNG_m}$ when using array_2.

6. CONCLUSIONS AND FURTHER WORK

In this work the effect of regularization on the appraisal of binaural reproduction was investigated. Firstly, we introduced an alternative definition of a WNG-criterion, which is better suited to re-synthesize HRTFs using microphone arrays.
Secondly, the evaluation of the perceived sensor noise (if noticeable) seems to improve considerably with increasing $\mathrm{WNG_m}$, whereas the explicit presence of sensor noise (primarily at lower $\mathrm{WNG_m}$ with array_1) does not consistently seem to deteriorate the overall performance. This latter observation may be due to the chosen test paradigm - it is conceivable that noise is more disturbing in other scenarios, e.g. when listening to music recordings. Furthermore, the higher sensor noise of array_1 seems also to have caused worse evaluations with regard to localization, coloration and overall performance for $\mathrm{WNG_m} \gtrsim -3$ dB. This phenomenon may be explained by the empirically higher Lagrangian multipliers μ that were required for array_1 to comply with a fixed $\mathrm{WNG_m}$ (cf. section 5.3).
The best compromise with regard to all assessed aspects and the associated robustness can be found at $\mathrm{WNG_m}$ of -6 dB and -3 dB for array_1 and at the highest of the tested $\mathrm{WNG_m}$ of 0 dB for array_2.
In general, the obtained evaluations confirm the validity of re-synthesizing HRTFs using microphone arrays in conjunction with individually suited $\mathrm{WNG_m}$. There is still room for improvement for the calculation and regularization of the filter coefficients, especially with regard to spectral coloration. Thus, one next step may be to elaborate a more appropriate and frequency-dependent regularization method.

7. ACKNOWLEDGMENTS

This project was partially funded by Bundesministerium für Bildung und Forschung under grant no. 17080X10, by Akustik Technologie Göttingen and by the Cluster of Excellence 1077 "Hearing4All", funded by the German Research Foundation (DFG).

8. REFERENCES

[1] V. Mellert and N. Tohtuyeva, "Multimicrophone arrangement as a substitute for dummy-head recording technique," in *In Proc. 137th ASA Meeting*, 1997, p. 3117.

[2] Y. Kahana, P.A. Nelson, O. Kirkeby, and H. Hamada, "A multiple microphone recording technique for the generation of virtual acoustic images," *The Journal of the Acoustical Society of America*, vol. 105, no. 3, pp. 1503–1516, 1999.

[3] J. Atkins, "Robust beamforming and steering of arbitrary-beam patterns using spherical arrays," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, New Paltz, NY, October 16-19 2011, pp. 237–240.

[4] E. Rasumow, M. Blau, M. Hansen, S. Doclo, S. van de Par, V. Mellert, and D. Püschel, "Robustness of virtual artificial head topologies with respect to microphone positioning errors," in *Proc. Forum Acusticum, Aalborg*, Aalborg, 2011, pp. 2251–2256.

[5] E. Rasumow, M. Blau, S. Doclo, M. Hansen, S. Van de Par, D. Püschel, and V. Mellert, "Least squares versus non-linear cost functions for a vitual artificial head," in *Proceedings of Meetings on Acoustics*. 2013, vol. 19, pp. –, ASA.

[6] D. N. Zotkin, R. Duraiswami, and N.A Gumerov, "Regularized hrtf fitting using spherical harmonics," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, New Paltz, NY, October 18-21 2009, pp. 257–260.

[7] Cesar D. Salvador Castaneda, Shuichi Sakamoto, Jorge A. Trevino Lopez, Junfeng Li, Yonghong Yan, and Yoiti Suzuki, "Accuracy of head-related transfer functions synthesized with spherical microphone arrays," in *Proceedings of Meetings on Acoustics*. 2013, vol. 19, pp. –, ASA.

[8] Shuichi Sakamoto, Satoshi Hongo, Takuma Okamoto, Yukio Iwaya, and Yoit Suzuki, "Improvement of accuracy of three-dimensional sound space synthesized by real-time "senzi", a sound space information acquisition system using spherical array with numerous microphones," in *Proceedings of Meetings on Acoustics*. 2013, vol. 19, pp. –, ASA.

[9] S. Doclo and M. Moonen, "Design of broadband beamformers robust against gain and phase errors in the microphone array characteristics," *IEEE TRANSACTIONS ON SIGNAL PROCESSING*, vol. 51, no. 10, pp. 2511–2526, October 2003.

[10] K.U. Simmer, J. Bitzer, and C. Marro, "Post-filtering techniques," in *Microphone Arrays*, Michael Brandstein and Darren Ward, Eds., Digital Signal Processing, pp. 39–60. Springer Berlin Heidelberg, Berlin, Heidelberg, New York, May 2001.

[11] J. Bitzer and K.U. Simmer, "Superdirective microphone arrays," in *Microphone Arrays*, Michael Brandstein and Darren Ward, Eds., Digital Signal Processing, pp. 19–37. Springer Berlin Heidelberg, Berlin, Heidelberg, New York, May 2001.

[12] E. Mabande, A. Schad, and W. Kellermann, "Design of robust superdirective beamformers as a convex optimization problem," in *Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference on*, April 2009, pp. 77–80.

[13] Jian Li, Petre Stoica, and Zhisong Wang, "On robust capon beamforming and diagonal loading," *Signal Processing, IEEE Transactions on*, vol. 51, no. 7, pp. 1702–1715, July 2003.

[14] Ole Kirkeby and Philip A. Nelson, "Digital filter design for inversion problems in sound reproduction," *J. Audio Eng. Soc*, vol. 47, no. 7/8, pp. 583–595, 1999.

[15] D. Hammershøi and H. Møller, "Sound transmission to and within the human ear canal.," *Journal of the Acoustical Society of America*, vol. 100, no. 1, pp. 408–427, 1996.

[16] E. Rasumow, M. Blau, M. Hansen, S. Doclo, S. van de Par, D. Püschel, and V. Mellert, "Smoothing head-related transfer functions for a virtual artificial head," in *Acoustics 2012*, Nantes, France, April 2012, pp. 1019–1024.

[17] E. Rasumow, M. Blau, M. Hansen, S. van de Par, S. Doclo, V. Mellert, and D. Püschel, "Smoothing individual head-related transfer functions in the frequency and spatial domains," *Journal of the Acoustical Society of America*, 2014, accepted for publication.

Proc. of the EAA Joint Symposium on Auralization and Ambisonics, Berlin, Germany, 3-5 April 2014

3D REPRODUCTION OF ROOM AURALIZATIONS BY COMBINING INTENSITY PANNING, CROSSTALK CANCELLATION AND AMBISONICS

Sönke Pelzer, Bruno Masiero, Michael Vorländer
Institute of Technical Acoustics,
RWTH Aachen University,
Aachen, Germany
{spe,mvo}@akustik.rwth-aachen.de

ABSTRACT

Popular room acoustic simulations use hybrid models for precise calculation of the early specular reflections and stochastic algorithms for the late diffuse decay. Splitting the impulse response into early and late parts is also psychoacoustically reasonable. The early part is responsible for the localization and the spatial and spectral perception of sources, which makes the correct reproduction of its time-frequency structure important. In contrast the later part is responsible for the sense of spaciousness and envelopment, properties related to the room and its diffuse decay.

Nevertheless, in auralization systems the reproduction of the whole impulse response is done through the same reproduction system and method, even though there are systems better suited to coherent reproduction (important for the early arrivals of an impulse response) and others better suited for the reproduction of incoherent fields (the reverberant tail of an impulse response).

A hybrid approach is presented which uses one common loudspeaker system for the simultaneous rendering of different reproduction methods. A method with strong localization cues such as binaural via crosstalk cancellation or VBAP is used for the direct sound and early reflections, while a method with higher immersion and envelopment such as Ambisonics is used for the diffuse decay.

1. INTRODUCTION

The challenge of generating high quality artificial reverberation has been dealt with since the late 1950's with many studies and publications. Important insights about properties of a room impulse response were already derived, e.g. by Schroeder in 1954 [1]. Whilst this was mostly constrained to theoretical thoughts, early work in the field of artificial reverberation based on simple analogue feedback loops. The main goal in these times aimed at producing natural sounding reverberation [2]. Only after the introduction of the first computational algorithms for the estimation of real reflections, already known algorithms such as ray tracing (RT) and the image source method (ISM) were applied in acoustics by Krokstad in 1968 [3] and Allen and Berkley in 1979 [4]. From then on the focus shifted towards the replication of real and complex shaped rooms. Nevertheless, due to the high computational demand, it was not until 1984 that Borish [5] extended the popular image source model to arbitrary polyhedra. Until today, the combination of these two models in hybrid algorithms mark the state-of-the-art in room acoustics simulation and auralization techniques [6, 7, 8, 9], although more accurate approaches for the estimation of sound propagation in rooms are known. They base on Finite-Element-Methods (FEM), Boundary-Element-Methods (BEM) or Finite-Time-Differences (FDTD), but they suffer from high numerical demands on computation power and are thus hardly applicable for normal to larger rooms or broadband simulations including higher frequencies. Recent approaches used a combined wave and ray based simulation method, which calculates the lower end (e.g. below the Schroeder frequency) using the FEM [10]. Geometrically based simulations, such as the described RT or IS methods have, on the other hand, highly developed representatives that already realize real-time capabilities [11].

2. ROOMS ACOUSTICS, EARLY/LATE REFLECTIONS AND MIXING TIME

The room impulse response can be divided into an early part which is dominated by distinct strong early reflections and a late part that mainly consists of reflections which have been reflected and scattered several times, so that they thoroughly overlap due to increased reflection density over time and the broadening of the impulses with higher reflection orders. Many attempts have been made to define the transition time between these two parts on a physical basis, but recent conclusions show that physical mixing does not explain diffusion and does not define the moment when a sound field turns diffuse [12]. It is in question if a perfectly diffuse reverberation exists at all in a real room. As the motivation for the separation of the impulse response is based on a psychoacoustic effect, it can be concluded that the human auditory system is not able to distinguish single reflections anymore as from a certain reflection density, a consensus in literature [13, 14]. Thus the transition time can still be determined in perceptual investigations, of which many have been conducted in the recent years. Unfortunately, most of them were restricted to only one room [13, 14], so that generalized conclusions cannot be drawn.

A detailed comprehensive overview of physical predictors for the estimation of the transition time as well as their evaluation on a perceptual basis can be found in a recent publication by Lindau [15]. The investigated predictors comprised model based ones (deriving the transition time from room parameters such as volume and mean free path length) as well as impulse response based ones (analyzing the time domain impulse response).

Shoebox shaped rooms usually have longer mixing times, due to their long unobstructed path length and regular shape. For these enclosures, Lindau found a transition time t_m, proportional to the mean free path length, with

$$t_m = 20V/S + 12 \quad [ms], \tag{1}$$

V being the room volume and S the room's surface area. Absorption and reverberation time were not found to have significant influence.

Regarding the IS model for prediction of early reflections, we find that the time range in an impulse response that is covered by a constant order of image sources is proportional to the mean free path length, just as the transition time t_m itself, as proposed by Lindau. This concludes to the necessary image source order O_{IS} being a constant factor between mean free travel time $t = 4V/cS$ (c: speed of sound) and transition time t_m:

$$O_{IS} \cdot \bar{t} = t_m \tag{2}$$

To estimate the necessary image source order, the additional 12 ms in the transition time formula will be neglected in favor of a full additional order of image sources, which is a valid approximation for even small rooms with at least 4 m of mean free path length. Including this simplification, the necessary image sources order can be estimated independently of reverberation time, volume or absorption to $O_{IS,min}$, with:

$$O_{IS,min} = \frac{t_{m-12}}{\bar{t}} + 1 \approx 2.7 \tag{3}$$

It can be concluded that for rooms, as selected by Lindau, which had shoebox shape and volumes in a wide range from $182m^3$ up to $8500m^3$, each with varied mean absorption, a general minimum IS order can be defined that results to three. After this third reflection, the sound field can be expected to be mostly mixing, uniform and isotropic, yielding a diffuse late reverberation. Similar observations were found by Kuttruff [16] when he analyzed the contributions of specular und diffuse energy in a room impulse response (RIR), as shown in Figure 1.

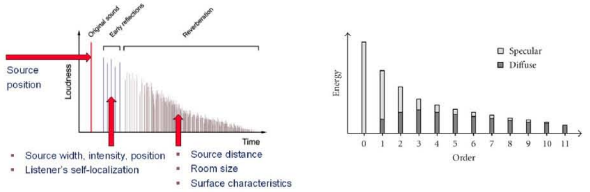


Figure 1: Left: *Perceptual and physical division of the room impulse response.* Right: *Relation of specularly and diffusely reflected sound in a typical room[16].*

Scattered reflections in the early part und all reflections after the image sources cut-off time (which should at least cover the mixing time) are then calculated using the ray tracing technique which calculates the temporal energy envelopes for each frequency band. The reflection modeling of image sources (IS) and RT are illustrated in Figure 2.

3. ROOM AURALIZATIONS USING SPATIAL 3D SOUND REPRODUCTION

To provide immersive auralizations the simulation results are processed so that they can be reproduced over headphones or loudspeakers. On the reproduction side it is important to remember the psychoacoustic motivation of the separation of components in the room impulse response. Early reflections and especially the direct sound have to be reproduced with highest precision in terms

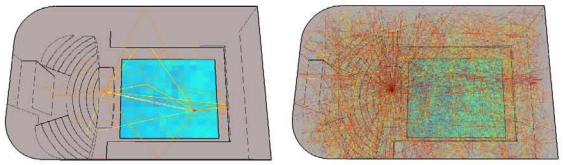


Figure 2: *Visualization of modeled reflections in a concert hall. Early reflections are constructed using the Image Source method (left) while late reverberation is modeled using ray tracing (right).*

of time and direction of arrival and frequency spectrum. Due to the precedence effect, the direct sound has a major influence on the localization of a source and the early reflections will affect the perceived source width. The reproduction system has to make sure that localization is as natural as possible, including exact compliance with frequency-dependent interaural level and time differences [17].

The point of full three-dimensionality is missed in many spatial reproduction techniques. A 3D reproduction should include not only horizontally distributed sources, but also the incidence from elevated angles and near field effects for sources that are close to the head of the listener [18]. Even large and expensive wave-field synthesis (WFS) systems mostly do not provide height information. More commonly used and more affordable systems such as vector-base amplitude panning (VBAP) and Ambisonics can theoretically reproduce elevated sources, but there are only few implementations that support realistic distance perception. VBAP has no support for close-by sources and Ambisonics only in near-field compensated higher-order ambisonics (NFC-HOA) setups [19].

Regarding this, binaural technology has a lot of advantages in 3D rendering, being very close to the way how the human ear perceives sound in nature. But as a major disadvantage it is difficult to reproduce binaural cues using loudspeakers. Using headphones on the other hand is not only problematic in terms of comfort and externalization, but also usually not able to impart the feeling of envelopment in diffuse sound fields. Additional problems such as the necessity to compensate for individual headphone transfer functions accrue.

A popular method to reproduce binaural signals is the crosstalk cancellation (CTC) [20], also called transaural in some publications. It uses a regular loudspeaker system, with only two speakers required, and takes advantage of wave interference to achieve a sufficient channel separation between the left and right ear of the listener. The main drawback of this system is the requirement to accurately know the current position of the user, which is typically solved using a tracking system and continuous adaption of the CTC filters [21]. Thus, this technique is often found in virtual reality systems, when the user is already tracked for interaction or 3D visualization [22].

Guastavino et al. [23] compared different reproduction techniques (CTC, Ambisonics, Panning) and came to similar results as described above and summarized in Table 1 (with additional comments by this author). It can be concluded that the reproduction method must also account for the psychoacoustics that define our hearing in rooms.

Table 1: *Comparison of different reproduction techniques, as published by Guastavino [23] with additional comments.*

Method	Advantages	Drawbacks
Binaural CTC	Precise localization, good readability, near field sources	Poor realism, lack of immersion/envelopment, needs individual HRTF
Ambisonics	Strong immersion and envelopment	Poor localization/readability
Stereo Panning	Precise localization	Lack of immersion/envelopment

4. HYBRID REPRODUCTION SYSTEMS

The idea to combine different systems for a separated reproduction of direct sound and reverberation was first mentioned in the Ambiophonics group in the early 1980s, mainly supported by Glasgal, Farina and Miller [24]. Their approach proposed a crosstalk canceled stereo-dipole playback for a wider stereo image and optional additional ambience speakers fed by the original signal convolved with an IR of a hall or similar reverberant space. The idea mainly aimed at an advanced reproduction of commercially available stereo recordings that were performed with certain popular microphone arrangements, such as ORTF or M/S, but the group also proposed their own microphone methodology and called it Ambiophone: two head-spaced omnidirectional microphones with a baffle behind them to muffle room reflections from non-frontal directions. Farina combined the stereo-dipole technique then with Ambisonics and had the chance to convolve his recordings with Ambisonics impulse responses of the hall where the recordings were actually made. The application of Ambiophonics can mainly be seen in the enhancement of stereo or 5.1 recordings, but the optional ambience channels have to be seen more as an artificial effect due to the fact that general recordings do not come with spatial impulse responses of the recording venue.

It was not before 2010 that Favrot proposed to apply the idea of hybrid reproduction that is matched to the events in a RIR to room acoustics prediction models which can generate spatial IRs for existing or virtual halls [25]. He used a variable Ambisonics order for the early and late part of the RIR to benefit from reduced computation load for late reverberation and better localization of the direct sound.

In this present contribution a combined hybrid system is introduced that uses one common loudspeaker system to play a CTC and an Ambisonics signal at the same time. The binaural signal will ensure high detail of temporal and spectral features of the direct sound and early reflections, while the Ambisonics signal is used to produce a spacious and enveloping diffuse sound field. The poor localization abilities of Ambisonics are published in a variety of studies [23, 26], and the poor immersion of binaural or transaural reproduction is documented as well [23]. Both observations clearly motivate the hybrid approach where binaural signals are used for the direct sound and early reflections and Ambisonics for the late decay.

Table 1 shows how the Pros and Cons of these two technologies are close to being perfectly complementary. The benefit and effort of binaural signals that use individual head-related transfer function (HRTF) are recently discussed. An investigation by Majdak found that a mismatch and lack of individualization substantially degraded the localization performance of targets placed outside of the loudspeaker span and behind the listeners, showing the relevance of individualized CTC systems for those targets [27].

The earlier introduced transition time is perfectly qualified to define the crossover between the two reproduction systems, with the same motivation as for the simulation. Therefore the CTC is used to reproduce the direct sound and specular reflections up the order of 3. Further reflection paths and all scattered reflections are fed into the Ambisonics engine. The presented idea is not meant to replace any cinema or public address system, due to the fact that the CTC is a single-user experience. It is more aimed at sophisticated room acoustics simulation and reproduction in virtual acoustics applications, such as virtual concert hall prototyping or fully immersive virtual environments [22].

4.1. Binaural Synthesis

Binaural filters are generated by attenuating each audible image source according to the distance law for spherical sources. The absorption coefficients of all walls in the reflection path are combined to a spectral filter which is then convolved with the source directivity. The last step of this filter chain adds the spatial information by including the HRTF data for the correct sight angle of the image source.

Virtual sound sources closer than 2 m need appropriate HRTF data that is measured in the near field (0.2, 0.3, 0.4, 0.5, 0.75, 1.0, 2.0 meters) [21]. If no such near field data is available, a range extrapolation should be applied, as proposed by Pollow [28]. If a NFC-HOA decoder is available, the range extrapolation can be implemented by interpreting a set of fixed-range HRTF as a virtual loudspeaker array [29].

To be able to compare also discretely working 3D reproduction systems, the binaural IR can also be extended to comprise all late reflections, so that the full room impulse response is ready for playback trough a CTC system.

4.2. Crosstalk Cancellation

A loudspeaker-based binaural reproduction chain starts with a binaural signal that can be either recorded using an artificial head or, in case of the presented method, simulated and synthesized by convolution of a HRTF or binaural RIR with an anechoic signal. A crosstalk cancellation filter network makes sure that the original binaural signal arrives at the listeners eardrums. Ideally, the CTC filters have to be constantly adjusted to the listener's head position and rotation. Combined with a dynamic binaural synthesis, a dynamic CTC allows a realistic spatial reproduction with only few loudspeakers. Dynamically adjusting CTC filters and binaural IRs in real-time requires considerable system complexity and low-latency convolution, both available since a few years [21]. The tracking devices often base on electromagnetic or optic input, but current developments aim at contact-free 6-degrees-of-freedom tracking by using infrared depth maps or video-based face detection.

When combined with an Ambisonics or other reproduction setup that already offers a multiple loudspeaker installation, it is possible to select two speakers of this setup which will serve the best possible channel separation. In a dynamically tracked system, this loudspeaker pair can be continuously exchanged dependent on the user's head position, without noticeable switching artifacts, as shown by Lentz [21]. Equation (4) describes the CTC as a closed-form solution, with $Z_{L/R}$ denoting the perceived signal:

$$\begin{bmatrix} Z_L \\ Z_R \end{bmatrix} = \begin{bmatrix} H_{LL} & H_{LR} \\ H_{RL} & H_{RR} \end{bmatrix} \begin{bmatrix} Y_L \\ Y_R \end{bmatrix} = \mathbf{Hy} = \mathbf{z} \tag{4}$$

The filters for the CTC are placed prior to the loudspeakers, so that $y = C \cdot x$. The transfer function of the complete system is given in matrix form as

$$\mathbf{z} = \mathbf{H} \cdot \mathbf{C} \cdot \mathbf{x} \tag{5}$$

For a binaural reproduction the output z should be equal to the input x apart from a time delay. Thus, the following equation has to be valid:

$$\mathbf{H} \cdot \mathbf{C} = \mathbf{e}^{-\mathbf{j}\boldsymbol{\Delta}} \cdot \mathbf{I} \tag{6}$$

with I being the identity matrix. The transfer matrix with the crosstalk cancellation filters C can easily be obtained by means of a pseudo-inverse of the transfer matrix H, resulting in:

$$\mathbf{C} = \mathbf{e}^{-\mathbf{j}\boldsymbol{\Delta}} \cdot \mathbf{H}^{+} \tag{7}$$

The closed-form solution according to equation (7) is the exact solution for the entire crosstalk cancellation. It requires, however, infinitely long filters that are also prone to have stability problems. The later issue can be dealt with through a regularized matrix inversion approach.

The regularization applies a constraint at the maximum gain allowed to the filters and can be expressed as follows:

$$\mathbf{C} = \mathbf{e}^{-j\Delta} \cdot (\mathbf{H}^{H}\mathbf{H} + \beta(f)\mathbf{I})^{-1} \cdot \mathbf{H}^{H} \tag{8}$$

This approach has no requirement on the matrix $\mathbf{H}$ to be square, meaning that more than two loudspeakers could be simultaneously used to achieve improved channel separation [30].

4.3. Higher-Order Ambisonics

The Ambisonics technique was initially designed in the early 1970s to perform spatial recordings and multi-channel broadcasting [31]. The known recording method of intensity stereophony with two perpendicularly superposed cardoid microphones (XY-arrangement) was upgraded by the use of an extra figure-of-eight microphone perpendicular to the other two microphones – note that the XY-arrangement with two cardioids can be substituted by two figure-of-eight and an omnidirectional microphone. This configuration corresponds already to the 0th and 1st spherical harmonics (SH) orders. Moreover, the original formulation of 1st order Ambisonics can be expanded to higher spherical harmonics orders, the so called higher-order ambisonics (HOA). This improves the usually imprecise localization of only 1st order reproduction at the cost of a more complex recording and reproduction system.

An Ambisonics microphone with three figure-of-eight microphones plus one omnidirectional microphone at the same position is unpractical. However, a set-up with four omnidirectional microphones on the faces of a tetrahedron can be used instead by later transforming the signals into the desired omnidirectional and figure-of-eight patterns using spherical harmonics transformation. The microphone signals are called A-format while the transformed signals are called B-format. The B-format signals can be independently stored or broadcasted and for playback they are adequately decoded into the G-format which is directly fed into the speaker set-up available for reproduction. The B-format guarantees storage and transmission of spatial audio data, independent of the decoding stage. The decoding step is then only dependent on the available loudspeaker setup, which has to be dimensioned to fulfill the requirements of the Ambisonics order N, i.e. the number of loudspeakers L has to be at least:

$$L \approx (N+1)^2 \tag{9}$$

In the proposed hybrid system Ambisonics is only used for late reflections, therefore the usual implementation of plane wave sources would be sufficient and near-field compensation (NFC) [19] is not essential. However, to include comparisons of direct sound rendering using the different methods, also a NFC decoder has been implemented.

After the design of a CAD room model and its parametrization, including material properties, source positions/directivities and receiver positions/HRTF, the IS model will return the positions and spectra of audible image sources and the ray tracer returns spatially discretized time-frequency energy histograms. To auralize the virtual scene, this information can now be translated into actual impulse responses. As proposed, the early reflections part is rendered into a binaural IR, while the scattered and late reflections are used to build an Ambisonics B-format IR.

4.3.1. Generation of Ambisonics B-format Impulse Responses

The late reverberation is predicted using a ray tracer. Thus, the simulation result is a data structure that contains the amount of energy that is arriving from a certain direction in a certain time interval in a certain frequency band, as shown in Figure 3. The temporal, spectral and spatial domains are discretized, usually in accordance with the number of rays for the desired resolutions.

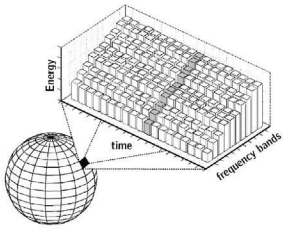


Figure 3: *Acoustic ray tracing results in a spatial data structure with time-frequency information of the energy of incident rays for each detection sphere.*

As the number of rays is meant to be kept a variable parameter and is usually chosen to a much lower number compared to the amount of real reflections in a room, the late reverberation is modeled by a synthetic sequence of Dirac pulses. If these pulses are arranged in accordance with the exponentially growing actual reflection densities over time, as derived by Kuttruff [16], then the whole sequence describes a Poisson process. The spectrum of this Poisson sequence is flat, so that no coloration occurs. In order to apply the temporal envelope of the ray tracing result for a certain frequency band, the noise sequence is filtered through an octave filter bank. Before the distinct pulses are smeared and overlapped by the filter bank, they are at first weighted by spherical harmonics. Therefore each single pulse is inserted in an own channel for each spherical harmonics order/degree with an amplitude according to the direction of arrival which is known by the ray tracing results. The band filtered Poisson sequences, which contain already temporally and spatially weighted pulses, are then superposed and result in the channels of a broadband B-format impulse response.

To enable comprehensive listening tests, it is also possible to render all image sources additionally into the B-format, so that the whole room impulse response (including direct sound and early reflections) can be played back through the Ambisonics system.

As for the IR synthesis algorithm, there is no limitation to the maximum SH order. In practice, 1st order reproduction can be sufficient in a hybrid system, while the cues that are important for localization are covered by the CTC. Anyway, if the late decay is not spatially homogeneous, e.g. in case of late echos from certain directions, a higher order encoding will improve the localization of late reflections.

In preliminary informal listening tests, the 2nd order signals were judged better than 1st order signals, even when only applied for late reflections during combined hybrid CTC/Ambisonics playback. Especially for cases when unusual room shapes should be simulated, such as long L-shaped corridors, it will become important to have spatially coded late reverberation. Problems as reported by other authors [32, 33], who encountered phenomena such as coloration or inside-head localization due to the correlation of the loudspeaker feeds, were not noticed in the test trials with the presented method. However, the decision if late reverberation should be spatially coded and thus have correlated loudspeaker feeds or if uncorrelated noise should be used which assumes a perfectly diffuse sound field and misses out on any spatial cues is subject to upcoming listening tests in the near future.

4.4. Calibration

To be able to seamlessly mix the early and late part of the IR with different reproduction techniques it must be ensured that their levels are accurately adjusted. In an ideal case under free field conditions it is possible to calculate the resulting sound pressure levels at a single position for any of the presented techniques. Under real world conditions the perfect sweet spot does not exist because the signals are presented to a human listener with two ears. In case of Ambisonics and CTC the assumption of ideal interference of the signals in a single point (Ambisonics) or two ear drums (CTC) usually does not hold true.

Dependent on the actual speaker layout the position of a virtual source has also an impact on the resulting level. This is especially the case if the layout is not regular or if the mounting conditions of each speaker are not exactly the same (which is nearly impossible to achieve in a normal room).

Therefore it is hard to calculate the accurate binaural sound pressure levels. To equalize the levels as best as possible without any knowledge about the virtual scene or real listening room, an equal distribution of virtual sources on a sphere (>900 sources) was used. The listening room with installed loudspeaker system is then measured or simulated and the loudspeaker impulse responses are used for CTC, HOA and VBAP decoding. To prepare signals for an unknown listening room, the loudspeaker IRs can be simulated for free field conditions. However, the impact of the listening room on the final levels and the calibration between the different formats has not been analyzed yet. The author assumes that the impact is different for CTC compared to Ambisonics or VBAP.

A general investigation on the impact of the listening room on loudspeaker reproduction of auralizations that include reverberation was performed by the author and will be published [34].

5. LOCALIZATION PERFORMANCE LISTENING TESTS

The accuracy of localization of virtual sources was measured for different reproduction methods. A listening test was conducted in a fully anechoic room with a 24-channel loudspeaker array, as shown in Figure 4). The loudspeakers were arranged in three layers with elevations of 0° and $\pm 30^\circ$ and an azimuth angle of 45° between each loudspeaker starting with a frontal direction of 0° azimuth.

Figure 4: *Listening tests were conducted in an anechoic chamber equipped with 24 loudspeakers for spatial reproduction.*

Figure 5: *Tracked head-mounted display and virtual environment for accurate pointing.*

A tracked head-mounted display (HMD, Oculus Rift) provided an accurate and bias-free pointing method [35]. A three-dimensional virtual sphere was rendered with a grid in 15° resolution and reference lines for horizontal and median planes, as shown in Figure 5. The listener's current view direction and head orientation/rotation were shown. In a training phase only real loudspeakers were driven with pink noise and the HMD displayed the actual source position. Averaged over all subjects an average pointing accuracy of 0.3° was measured.

Five reproduction methods were tested. Three pure implementations of CTC, VBAP and 4th-order Ambisonics and two hybrid variants using CTC or VBAP for the early party and 4th-order Ambisonics for the late reverberation. For the CTC the HRTFs of the artificial head *Fabian*[36] were used and 2 loudspeakers at an elevation of 0° and an azimuth of $\pm 45^\circ$. These HRTFs were measured with a source distance of 1.7m and therefore matching the loudspeaker distance in the test chamber. Fourth order Ambisonics was used with plane wave max-$|r_E|$ Ambisonics decoding.

As virtual room a model of the Concergebouw in Amsterdam was used to provide a realistic environment. The receiver was placed over the first rows. Four sources were presented at a distance of 5.5m (critical distance in this room model) at positions

shown in Figure 6. The positions were limited to the front direction and to an elevation of $\pm 30^\circ$ to ensure a valid reproduction for VBAP and avoid the subjects having to turn around. Each position was repeated three times while the order of all stimuli was randomized.

A sound file was only played as long as the subjects were looking directly in front. A deviation of more than 2° would pause the playback immediately. The samples could be repeated as often as desired.

In total 18 subjects participated in the test with an average localization accuracy of 16°. The deviation was calculated as the distance of the cone of confusion of the presented source and the chosen direction. Individual results of all subjects are shown in Figure 7 (left). Between the presented five systems no significant differences were found in a one-way ANOVA test, as shown on the right hand side of Figure 7. The ANOVA yields a significant main effect of source position (F(3,51)=6.695; MSE=0.222; $p < 0.001$; η_p^2 =0.283) between position 4 and all other positions. A two-way ANOVA revealed a significant main effect of system for positions 2 (F(4,64)=9.463; MSE=0.075; $p < 0.001$; η_p^2=0.372) and 4 (F(4,64)=11.538; MSE=0.019; $p < 0.001$; η_p^2=0.419), as shown in Figure 8. It can be concluded that VBAP and HOA have a higher dependency on the source position than the CTC.

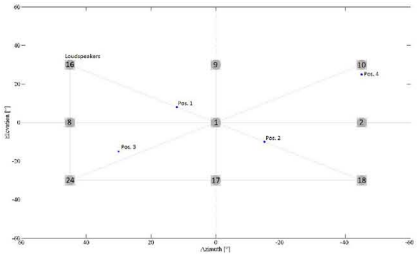


Figure 6: *Presented 4 source positions in the listening test. Frontal 9 loudspeakers are shown for reference.*

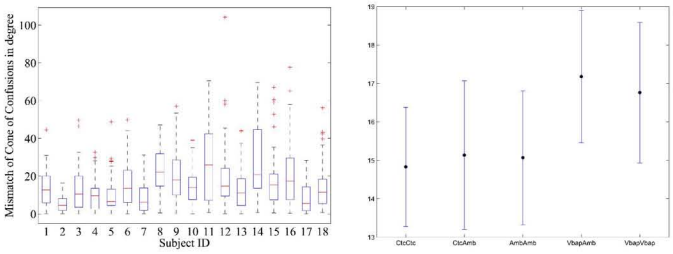


Figure 7: Left: *Results of the localization perfor-mance tests for the 18 individual subjects.* Right: *Results of the ANOVA for different reproduction systems (CTC, CTC+HOA, HOA, VBAP+HOA, VBAP). No significant differences were found in localization performance.*

6. CONCLUSIONS

A method was presented that combines different loudspeaker-based reproduction methods (such as CTC, Ambisonics or VBAP) to auralize a sound field. The sound field can consist of one or more sound sources and all reflections of these sources that bounce off

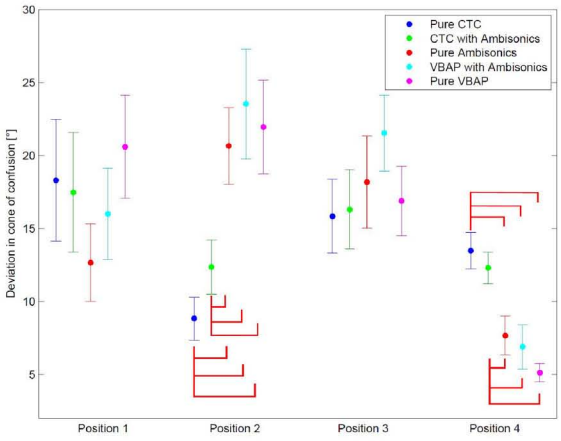


Figure 8: *Results of two-way ANOVA grouped by presented source position. The high impact of the presented source position is visible. Red brackets indicate significant differences.*

walls in the virtual scene. The simulation of the sound field including all room acoustics reflections is done using the RAVEN framework. For the hybrid auralization, the room impulse response is divided into three parts (direct sound, early reflections and late reverberation), according to findings from psychoacoustic research. This division also suits the algorithms of geometrical acoustics that are commonly used in room acoustics simulations (image sources and ray tracing).

The hybrid approach can take advantages of the individual strengths of each reproduction method. Strong localization cues are necessary for direct sound rendering. The late diffuse sound field should be rendered using immersive reproduction methods. Both signals are calculated using RAVEN and are played back simultaneously through the same loudspeaker setup. To enable a seamless transition the average loudness of the different systems has to be calibrated accurately, which has to be done for each individual loudspeaker setup.

The moment of transition from the early to the late part of the impulse response is defined by the mixing time. It was shown that in typical cases after three reflection orders the sound-field can be expected to be mixing and diffuse. Then the renderer can switch from a method with strong localization to a method with high envelopment. The aim is to render a realistic and natural sounding high quality auralization of spatial sound with reverberation.

With the used loudspeaker setup (24-channel array) none of the tested systems provided an overall superior localization performance than the other systems. However, the binaural CTC provided a more homogeneous localization accuracy across different source positions. This behavior is typically preferred, especially for scenes with moving sources, making this technique suitable for the early part of the impulse response. A test for the immersiveness of different systems has to be designed and conducted in further research, to find an optimal method for the reproduction of late reverberation.

7. REFERENCES

[1] M.R Schroeder, "Die statistischen parameter der frequenzkurven von großen räumen," *Acustica 4*, vol. 4, pp. 594–

600, 1954.

[2] M.R Schroeder, "Natural sounding artificial reverberation," *13th AES Convention*, 1961.

[3] S. S. A. Krokstad and S. Sorsdal, "Calculating the acoustical room response by the use of a ray tracing technique," *J. Sound and Vibration*, vol. 8, pp. 118–125, 1968.

[4] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *J. Acoust. Soc. America*, vol. 65:943, 1979.

[5] J. Borish, "Extension of the image model to arbitrary polyhedra," *J. Acoust. Soc. America*, vol. 75, pp. 1827–1836, 1984.

[6] M. Vorländer, *Auralization: Fundamentals of Acoustics, Modelling, Simulation, Algorithms and Acoustic Virtual Reality*, RWTHedition Series. Springer, 2011.

[7] G. M. Naylor, "Odeon - another hybrid room acoustical model," *Applied Acoustics*, vol. 38:131, 1993.

[8] CATT-Acoustic, *http://www.catt.se*.

[9] W. Ahnert and R. Feistel, "Ears auralization software," *J. Audio Eng. Soc. Vol.*, vol. 41 (11), pp. 897–904, 1993.

[10] S. Pelzer, M. Aretz, and M. Vorländer, "Quality assessment of room acoustic simulation tools by comparing binaural measurements and simulations in an optimized test scenario," *Acta acustica united with Acustica*, vol. 97, no. S1, pp. 102–103, 2011.

[11] D. Schröder, *Physically Based Real-time Auralization of Interactive Virtual Environments*, Ph.D. thesis, RWTH Aachen University, 2011.

[12] J.-D. Polack, "Is mixing the source of diffusion?," *J. Acoust. Soc. Am.*, vol. 129(4), pp. 2502–2502, 2011.

[13] A. Reilly, D. McGrath, and B.-I. Dalenbäck, "Using auralisation for creating animated 3-d sound fields across multiple speakers," *Proc. 99th AES Conv., New York*, vol. preprint no. 4127, 1995.

[14] D. Meesawat, K; Hammershøi, "The time when the reverberant tail in binaural room impulse response begins," *Proc. 115th AES Conv., New York*, vol. preprint no. 5859, 2003.

[15] A. Lindau, "Perceptual evaluation of physical predictors of the mixing time in binaural room impulse responses," *Proc. AES 128th Conv., London, UK*, 2010.

[16] H. Kuttruff, *Room Acoustics*, 4th ed., New York: Routledge Chapman & Hall, 2000.

[17] A. Avni and B. Rafaely, "Sound localization in a sound field represented by spherical harmonics," *Proc. 2nd Internat. Symposium on Ambisonics and Spherical Acoustics, Paris, France*, 2010.

[18] T. Lentz, D. Schröder, M. Vorländer, and I. Assenmacher, "Virtual reality system with integrated sound field simulation and reproduction," *EURASIP journal on advances in signal processing*, vol. 2007, pp. 70540, 19 S., 2007.

[19] J. Daniel, "Spatial sound encoding including near field effect: Introducing distance coding filters and a viable, new ambisonic format," *AES 23rd Internat. Conf., Copenhagen, Denmark*, 2003.

[20] B. B. Bauer, "Stereophonic earphones and binaural loudspeakers," *J. Audio Eng. Soc.*, vol. 9(2), pp. 148–151, 1961.

[21] T. Lentz, *Binaural technology for virtual reality*, Ph.D. thesis, RWTH Aachen University, 2011.

[22] D. Schröder, F. Wefers, S. Pelzer, D. Rausch, M. Vorländer, and T. Kuhlen, "Virtual reality system at rwth aachen university," in *Proc. ICA 2010, 20th Internat. Congress on Acoustics, Sydney, Australia*. 2010, Australian Acoustical Society, NSW Division, 1 CD-ROM.

[23] C. Guastavino, V. Larcher, G. Catusseau, and P. Boussard, "Spatial audio quality evaluation: comparing transaural, ambisonics and stereo," *Proc. 13th Internat. Conf. on Auditory Display, Montreal, Canada*, 2007.

[24] A. Farina, R. Glasgal, E. Armelloni, and A. Torger, "Ambiophonic principles for the recording and reproduction of surround sound for music," *Proc. AES 19th Internat. Conf.*, 2001.

[25] S. Favrot and J. M. Buchholz, "Lora: A loudspeaker-based room auralization system," *Acta Acustica united with Acustica*, vol. 96, pp. 364–375, 2010.

[26] V. Pulkki, "Evaluating spatial sound with binaural auditory model," *Proc. Internat. Computer Music Conference, Havana, Cuba*, pp. 73–76, 2001.

[27] P. Majdak, B. Masiero, and J. Fels, "Sound localization in individualized and non-individualized crosstalk cancellation systems," *J. Acoust. Soc. America*, vol. 133(4), pp. 2055–2068, 2013.

[28] M. Pollow, K.-V. Nguyen, O. Warusfel, T. Carpentier, M. Müller-Trapet, M. Vorländer, and M. Noisternig, "Calculation of head-related transfer functions forarbitraryfield points using spherical harmonics decomposition," *Acta acustica united with Acustica*, vol. 98(1), pp. 72–82, 2012.

[29] T. Musil M. Noisternig, A. Sontacchi and R. Höldrich, "A 3d ambisonic based binaural sound reproduction system," *AES 24th Internat. Conf. on Multichannel Audio, Banff, Canada*, 2003.

[30] B. Masiero, *Individualized binaural technology: measurement, equalization and perceptual evaluation*, Ph.D. thesis, RWTH Aachen University, 2012.

[31] M. A. Gerzon, "Periphony: With-height sound reproduction," *J. Audio Eng. Soc.*, vol. 21(1), pp. 2–10, 1972.

[32] A. Solvang, "Spectral impairment for two-dimensional higher order ambisonics," *J. Audio Eng. Soc.*, vol. 56, pp. 267–279, 2008.

[33] J. Daniel, *Representation de champs acoustiques, application a la transmission et a la reproduction de scenes sonores complexes dans un contexte multimedia (in french)*, Ph.D. thesis, Universite Paris 6, 2000.

[34] S. Pelzer and M. Vorländer, "Auralization of virtual rooms in real rooms using multichannel loudspeaker reproduction," *J. Acoust. Soc. America*, vol. 134, pp. 3985–3985, 2013.

[35] P. Majdak, M. J. Goupell, and B. Laback, "3-d localization of virtual sound sources: Effects of visual environment, pointing method, and training," *Attention, Perception, & Psychophysics*, vol. 72, no. 2, pp. 454–469, 2010.

[36] A. Lindau and S. Weinzierl, "Fabian-schnelle erfassung binauraler raumimpulsantworten in mehreren freiheitsgraden," *Fortschritte der Akustik 33.2: 633*, 2007.

Proceedings of the EAA Joint Symposium on Auralization and Ambisonics

In consideration of the remarkable intensity of research in the field Virtual Acoustics, including different areas such as sound field analysis and synthesis, spatial audio technologies, and room acoustical modeling and auralization, it seemed about time to organize a second international symposium following the model of the first EAA Auralization Symposium initiated in 2009 by the acoustics group of the former Helsinki University of Technology (now Aalto University). Additionally, research communities which are focused on different approaches to sound field synthesis such as Ambisonics or Wave Field Synthesis have, in the meantime, moved closer together by using increasingly consistent theoretical frameworks. Finally, the quality of virtual acoustic environments is often considered as a result of all processing stages mentioned above, increasing the need for discussions on consistent strategies for evaluation. Thus, it seemed appropriate to integrate two of the most relevant communities, i.e. to combine the 2nd International Auralization Symposium with the 5th International Symposium on Ambisonics and Spherical Acoustics. The Symposia on Ambisonics, initiated in 2009 by the Institute of Electronic Music and Acoustics of the University of Music and Performing Arts in Graz, were traditionally dedicated to problems of spherical sound field analysis and re-synthesis, strategies for the exchange of ambisonics-encoded audio material, and – more than other conferences in this area – the artistic application of spatial audio systems.

The publication at hand contains the official conference proceedings. It includes 29 manuscripts which have passed a 3-stage peer-review with a board of about 70 international reviewers involved in the process. Each contribution has already been published individually with a unique DOI on the DepositOnce digital repository of TU Berlin. Some conference contributions have been recommended for resubmission to Acta Acustica united with Acustica, to possibly appear in a Special Issue on Virtual Acoustics in late 2014. These are not published in this collection.

ISBN 978-3-7983-2704-7 (online)

Universitätsverlag der TU Berlin